UNIVERSITY OF HAWAIʻI AT MĀNOA

Integrative transcriptomic analysis of long intergenic non-coding RNAs in cancer

by

Travers H. Ching

A thesis submitted in partial fulfillment for the
degree of Doctor of Philosophy

Committee Members:
Lana X. Garmire
Ben Fogelgren
Maarit Tiirikainen
Brandon Yoza
Herbert Yu
Molecular Biosciences and Bioengineering

May 2017

Acknowledgements

Completion of this research would not have been possible without my advisor, Dr. Lana Garmire and my committee members, Dr. Ben Fogelgren, Dr. Maarit Tiirikainen, Dr. Brandon Yoza and Dr. Herbert Yu. In addition, I would like to thank Dr. Karolina Peplowska, Dr. Janos Molnar, and all of my fellow lab members.

Abstract

Long non-coding RNAs (lncRNA) are a relatively new and mysterious class of RNA molecules that are transcribed in eukaryotic cells. They are differentiated from mRNA transcripts in that they do not code for proteins and are much larger than small RNA species, such as microRNAs. Long intergenic non-coding RNAs (lincRNAs) are a subclass of lncRNAs that appear outside the boundaries of known genes. At the current state, little is definitively known about lincRNAs. LincRNAs have a diverse range of functions, such as providing molecular scaffolding for chromatin remodeling, acting as molecular sponges for microRNAs, or directly interacting with promoter and enhancer regions to promote or downregulate gene expression.

In humans, there are a huge number of lincRNA genes, more than the number of genes that are protein coding; it has been estimated that 80% of the human genome is transcribed, yet only 2-3% is translated. There is active debate in the field as to what proportion of those transcripts are biologically relevant, as the alternative is that some of those transcripts are meaningless noise, due to leaky RNA polymerases.

There is an increasing number of lincRNAs that are known to be functionally relevant to cancer such as XIST, MALAT1, HOTAIR and PCAT1. XIST generally acts to silence one copy of the X-chromosome in women; in breast cancer, it is found to be downregulated. HOTAIR, within the HOX locus, is deregulated in aggressive metastatic tumors. HOTAIR expression is increased in metastatic cancer and is a biomarker for poor prognosis. MALAT1 is enriched in the nucleus, regulates cell motility and is also implicated in metastasis. PCAT1 is implicated in disease progression in prostate cancer. However, the functions and mechanisms of most lincRNAs are not definitively known.

This dissertation focuses on elucidating the roles of lincRNAs in relation to cancer pathogenesis. The focus is on identifying lincRNA biomarkers in cancer and to further elucidate clinically relevant lincRNA mutations. Using bioinformatics and computational biology approaches to analyze lincRNA expression and mutation profiles, I will attempt to determine which lincRNAs

are relevant to tumorigenesis and progression and how mutation data correlates with expression and clinical phenotypes.

Information generated from this investigation will provide knowledge on the role of non-coding RNAs in the development and progression of cancer. It will also help to elucidate the application of machine learning methods to cancer and non-coding gene research domains. Most importantly, it will push forward the translational and clinical applications of lincRNAs as potential cancer biomarkers and therapeutic targets.

In chapter 1, I further explain the technical background relevant to the projects contained in this dissertation. Chapter 2 is a lincRNA review paper published in *BioData Mining*, focusing on the upcoming computational challenges related to lincRNA research. Chapter 3 is an analysis of RNA-Seq differential expression methods published in *RNA*; computational approaches in order to find upregulated or downregulated lincRNAs. Chapter 4 is an exploration of the expression landscape of lincRNA across 12 cancer types, published in *eBioMedicine*. Chapter 5 and 6 are applications of machine learning methods to high dimensional biological data. In Chapter 5, I explore a neural network-cox regression machine learning hybrid model, in order to predict patient survival, and to elucidate the biological pathways relevant to each patient. In chapter 6, I elucidate the somatic mutation landscape in lincRNAs across 12 cancer types. I quantify which molecular features are correlated with lincRNA mutation probability, and I show that these results could be used to provide more robust subtyping and clustering of tumor samples. Finally, in Chapter 7, I discuss what these research projects have accomplished in the grand scheme of the lincRNA research field, and explain what further work needs to be accomplished to follow up.

Specific aims

The goal of this work is to discover novel cancer biomarkers and to elucidate cancer biology through the application of machine learning methods using "omics" data, particularly on the expression and mutations of lincRNAs. I will apply classification and regression algorithms to elucidate novel biological insights in cancer biology. The roles of ncRNA in cancer will be addressed through completion of the following aims:

Specific aim 1: Use machine learning and statistical methods using RNA-Seq data to discover novel lincRNA biomarkers for cancer diagnosis or prognosis with preliminary functional evaluation through cell line experiments.

Specific aim 2: Find important relations between somatic mutations and expression in lincRNAs and how they relate to clinical features or subtypes.

Included manuscripts

1. Ching, T., Masaki, J., Weirather, J. & Garmire, L. X. Non-coding yet non-trivial: a review on the computational genomics of lincRNAs. *BioData mining* **8,** 44 (2015).
2. Ching, T., Huang, S. & Garmire, L. X. Power analysis and sample size estimation for RNA-Seq differential expression. *Rna* **20,** 1684–1696 (2014).
3. Ching, T., Peplowska, K., Huang, S., Zhu, X., Shen, Y., Molnar, J., Yu, H., Tiirikainen, M., Fogelgren, B., Fan, R. & Garmire, L. X. Pan-cancer analyses reveal long intergenic non-coding RNAs relevant to tumor diagnosis, subtyping and prognosis. *EBioMedicine* **7,** 62–72 (2016).
4. Ching, T., Zhu, X. & Garmire, L. X. Cox-nnet: a neural network extension to Cox regression. *(under review)* (2017).
5. Ching, T. & Garmire, L. X. Pan-cancer analysis of expressed single nucleotide variants in long intergenic non-coding RNA. *(under review)* (2017).

Contents

Chapter 1

Background

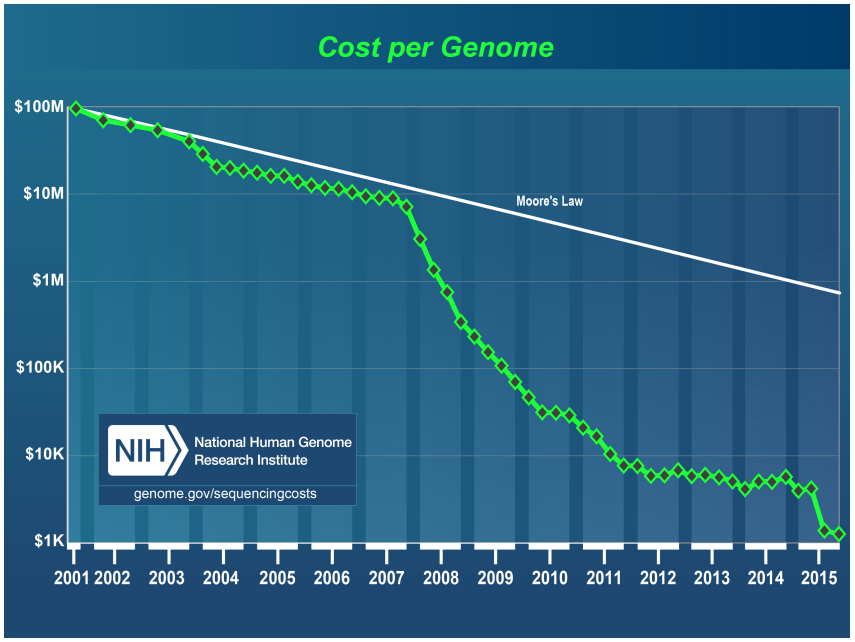
The application of machine learning methods for the meta-analysis of high-throughput genomics data can provide biological insights and discovery into the role that long intergenic non-coding RNA has with respect to cancer. In this chapter, I describe these terms and the aims and goals of this dissertation.

1.1 Next generation sequencing

Next generation sequencing (NGS) is a method that analyzes millions of short nucleic acids fragments (often 100 base pairs) by multiplexed parallel sequencing. In recent years NGS methods have significantly increased sequencing efficiency, producing faster results and also an exponential reduction in cost. In 2007, one human genome cost roughly $10 million to sequence, but now this cost has been reduced to $1000.

The most common sequencing platform is produced by the Illumina-Solexa Corporation. Before the two companies merged in 2007, Solexa had developed a new method of sequencing using fluorescent nucleotides rather than the traditional Sanger sequencing method of terminating nucleotides. The method relies on re-polymerization of short fragments using modified fluorescent nucleotides.

In the first step of the process, library preparation is performed by cutting genomic DNA (or cDNA) into 300-600 base pair fragments and ligated at both free ends using adapters. The ligated fragments are hybridized and bound to the surface of a flow cell. The flow cell also contain oligonucleotide primers used to initiate polymerization. Each fragment is amplified while fixed to the flow cell surface. Thermocycling (typically a 100 cycles) is performed using



Cost of sequencing by year (`http://genome.gov/sequencingcosts`)

tagged fluorescent nucleotides. Each flow cell can contain millions of hybridized fragments that are sequenced in parallel, resulting in massive data generation in a short time.

There are many applications to next generation sequencing. Two of the most common applications are DNA sequencing and RNA sequencing.

1.1.1 DNA sequencing

In the context of a human genome, DNA sequencing is often used for the purpose of discovering and individual's genetic alleles. DNA is typically extracted from a tissue sample and pre-amplified for library construction. Due to the size of the human genome, traditionally, only the exonic regions are analyzed, e.g., through probe specific amplification primers, or other similar mechanisms (termed Exome Sequencing or Exome-Seq). However, whole genome sequencing (synonymously called shotgun sequencing) has become more common allowing for the determination and analysis of previously unstudied genetic elements.

If the genome sequence is unknown, small nucleotide fragments must be assembled into progressively larger contigs for construction and determination of the full genome (i.e., de novo sequencing). When the reference genome is known short reads can be aligned against a genomic reference allowing for individual variation to be studied, including somatic tumor mutation determination.

The best known population scale Whole Genome Sequence (WGS) study in cancer research is The Cancer Genome Atlas (TCGA) project. TCGA began in 2005 with the goal of determining genome relationships for three different tumor types (glioblastoma, lung cancer and ovarian cancer). In 2009, TCGA expanded to include the determination of genetic elements associated with 20-25 different tumor types. Ordinarily, genomes must have roughly 10x sequencing coverage in order to retrieve genetic variants and alleles. In addition to the difficulties associated with sequencing a large genome, cancer identification in humans require greater sequencing coverage, especially for identification of rare somatic mutations and alleles, which may only occur in a small fraction of cells.

1.1.2 RNA sequencing (RNA-Seq)

RNA sequencing or RNA-Seq is the RNA counterpart to exome sequencing and whole genome sequencing, utilizing RNA as template material instead of DNA. Whereas Exome-Seq and WGS aim to sequence the whole exome or whole genome respectively, RNA-Seq aims to sequence the whole transcriptome.

It is quickly replacing microarrays as the platform of choice for gene expression profiling, owing to greater sensitivity and lower noise levels. Ribosomal RNA is usually removed from the library preparation, through poly-A selection. This RNA is then converted to cDNA through reverse transcription.

In addition, RNA-Seq has also greatly augmented our understanding of mechanisms of alternative splicing and has led to the discovery of novel isoforms and novel genes. This includes the discovery of thousands of novel, robustly expressed long intergenic non-coding RNAs (lincRNAs).[1–4]. Machine learning methods in cancer genomics

1.2 Machine learning in cancer

Many machine learning algorithms, including Artificial Neural Networks, Support Vector Machines and Decision Trees have been used extensively for understanding cancer genomics by allowing for the development of predictive diagnosis and prognosis models and for therapeutic intervention. However, due to massive size of available digitized biological data, finding important signals within the large amount of biological and methodological noise can be challenging.

There are a handful of clinically related factors for cancer biology that researchers are interested in; 1) Prediction of disease risk, 2) diagnosis of disease, 3) prediction of survival or tumor

recurrence and 4) prediction of drug response. Traditionally, each individual biomarker has been tested for its relevance to a particular aspect of cancer biology (such as disease progression or tumor subtype) without regard for potential synergistic interactions between markers.

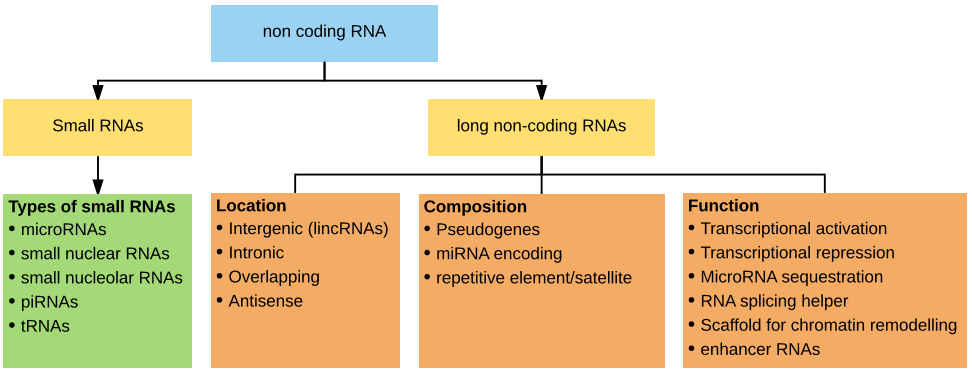
For example, genome-wide association studies have found that only 5% of the heritable risk for colorectal cancer. However, analysis of familial inheritance patterns have shown that heritability has a much more significant association of 35% [5]. The difference between the two examinations can be attributed to non-linear gene-gene interactions. Studies such as these suggest that cancer biology is clearly more complicated than the simple linear combination of markers. Many important indicators may be missed if a variety of possible interactions are not taken into consideration. E.g., interactions between mutations on the same gene or between different omics level (i.e., between mutations and the expression level of a gene). While many non-linear machine learning methods can take into account interaction terms (such as radial support vector machines), overfitting and multiple hypothesis testing are major problems. Furthermore, finding interaction terms in a machine learning framework requires greater sample size and interpretation is difficult and may often yield inconclusive results.

1.3 Non-coding RNAs

Non-coding RNAs (ncRNAs) are a large superclass of RNA molecules, defined as RNAs that are non-ribosomal and are not translated into proteins. While there is some discrepancy with how researchers classify non-coding RNAs, generally they are divided into two sub-groups based on size: small RNAs and long non-coding RNAs.

1.3.1 Small RNAs

Small RNAs are less than 200 base pairs, and are generally better characterized compared with long non-coding RNAs. In all mammals, small RNAs are divided into at least five classes: 1) microRNAs - attenuate translation of mRNAs, 2) small nuclear RNAs - assist in splicing of mRNAs, 3) small nucleolar RNAs - assist in post-transcriptional modification of RNA such as methylation and pseudouridinylation, 4) piRNAs - suppression of retro-transposons, and 5) tRNAs - transfer amino acids for peptide synthesis. MicroRNAs, as an example, are 20-25 base pairs and when properly spliced, their function is to attenuate gene expression. MicroRNAs form an imprecise complementary RNA-protein hybrid with the RNA induced silencing complex (RISC). The RISC complex expresses an enzyme that silences gene expression by cleaving mRNA.



Types and defining characteristics of non-coding RNA

1.3.2 Long non-coding RNAs

Long non-coding RNAs (lncRNAs) are a much less understood class of RNA molecules having diverse functions and origins. LncRNAs are usually categorized by their location on the genome, relative to known genes [6]. lncRNAs may be located within intronic regions, transcriptionally overlapping coding genes, anti-sense to coding genes or intergenic regions (lincRNAs) [6].

LncRNAs are sometimes also categorized by their compositional content and putative function. LncRNA may have sequence content that is homologous to protein coding genes and are termed pseudogenes. LncRNAs may contain repetitive elements, such as tandem repeats, endogenous retroviruses or microsatellites. lncRNAs act as transcriptional co-activators (enhancer RNAs) or repressors that can sequester microRNAs and therefore downregulate microRNA function (competitive endogenous RNAs or ceRNAs). lncRNAs have even been discovered to act as scaffolds during chromatin remodeling [**Wang2001**, **Prensner2011a**].

1.3.3 LincRNAs

Finally, lincRNAs are a subset of lncRNAs and are defined as non-coding RNAs that are greater than 200 bp in length and arise from intergenic regions within the genome. lincRNAs are the focus of this dissertation. The majority of lincRNAs have only recently been discovered, through genome wide RNA-Seq studies. One of the first high throughput studies, termed the "Human Body Map project" [7] identified 9000 lincRNAs in the human genome through RNA sequencing of many different human tissues. Many more lincRNAs have since been discovered, and various projects have undertaken the task of providing a more comprehensive transcriptional annotation. For example, the Lncipedia database has annotated 63,000 long non-coding RNAs, of which approximately 30,000 are robustly transcribed lincRNAs.

It is understood that protein coding exons constitute only about 3% of the human genome [8]. There are large regions of the human genome that have no protein coding potential. Many of these "gene deserts" actually encode for long, "intergenic" non-coding RNAs (lincRNAs). Furthermore, the large majority of RNA transcripts have been recently determined to actually be non-coding. According to the Encyclopedia of DNA Elements (ENCODE) project, about 62% of the entire genome is transcribed into long (>200 base pairs) RNA sequences [8]. Given that 3% of the genome encodes protein coding exons, majority of transcripts longer than 200 bp are non-coding. Of these long non-coding RNAs, roughly one third come from intronic regions or overlap with protein-coding genes, whereas about two thirds comes from intergenic regions [8]. While the function of intronic and overlapping ncRNA can often be attributed to regulation or association with the host gene, lincRNAs have no proximal association with genes and are somewhat more mysterious. In fact, the functions of almost all lincRNAs are not definitively known.

Although lincRNAs are believed to not be translated into proteins, lincRNAs are transcribed by polymerase II and often have poly-A tails and 5' Methyl cap. There are several well-studied lincRNAs associated with cancer, such as MALAT1, HOTAIR and PCAT1 [9, 10]. HOTAIR is found within the HOX loci and is deregulated in aggressive metastatic tumors [10]. This results in overexpression and is a biomarker having poor clinical prognosis. MALAT1 is enriched in the nucleus, regulates cell motility and is also implicated in metastasis. PCAT1 is implicated in disease progression in prostate cancer. PCAT1 binds to the Polycomb Repressive Complex 2 (PRC2) which is a histone methyltransferase that alters histone code selectively in a sequence specific manner citePrensner2011.

There have been attempts to extrapolate the function of lincRNAs on a genome-wide scale based on the function of well-understood lincRNAs. The accuracy of these extrapolations have not been systematically assessed. Furthermore it is also debatable as to the percentage of lincRNAs that are functional as opposed to transcriptional "noise" – spurious transcripts that don't have biological functionality [11].

References

1. Kim, D. & Salzberg, S. L. TopHat-Fusion: an algorithm for discovery of novel fusion transcripts. *Genome Biology* **12.** ISSN: 1465-6906. doi:10.1186/gb-2011-12-8-r72 (2011).

2. Morin, R. D., Bainbridge, M., Fejes, A., Hirst, M., Krzywinski, M., Pugh, T. J., McDonald, H., Varhol, R., Jones, S. J. M. & Marra, M. A. Profiling the HeLa S3 transcriptome using randomly primed cDNA and massively parallel short-read sequencing. *Biotechniques* **45** (2008).

3. Trapnell, C., Pachter, L. & Salzberg, S. L. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* **25,** 1105–1111. ISSN: 1367-4803, 1460-2059 (2009).

4. Wang, Z., Gerstein, M. & Snyder, M. RNA-Seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics* **10,** 57–63. ISSN: 1471-0056 (2009).

5. Al-Tassan, N. A., Whiffin, N., Hosking, F. J., Palles, C., Farrington, S. M., Dobbins, S. E., Harris, R., Gorman, M., Tenesa, A. & Meyer, B. F. A new GWAS and meta-analysis with 1000Genomes imputation identifies novel risk variants for colorectal cancer. *Scientific reports* **5** (2015).

6. Ma, L., Bajic, V. B. & Zhang, Z. On the classification of long non-coding RNAs. *RNA biology* **10,** 924–933 (2013).

7. Cabili, M. N., Trapnell, C., Goff, L., Koziol, M., Tazon-Vega, B., Regev, A. & Rinn, J. L. Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. *Genes & development* **25,** 1915–1927 (2011).

8. ENCODE. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489,** 57–74. ISSN: 0028-0836 (2012).

9. Prensner, J. R., Iyer, M. K., Balbin, O. A., Dhanasekaran, S. M., Cao, Q., Brenner, J. C., Laxman, B., Asangani, I. A., Grasso, C. S. & Kominsky, H. D. Transcriptome sequencing across a prostate cancer cohort identifies PCAT-1, an unannotated lincRNA implicated in disease progression. *Nature biotechnology* **29,** 742–749. ISSN: 1087-0156 (2011).

10. Ulitsky, I. & Bartel, D. P. lincRNAs: genomics, evolution, and mechanisms. *Cell* **154,** 26–46. ISSN: 0092-8674 (2013).

11. Kowalczyk, M. S., Higgs, D. R. & Gingeras, T. R. Molecular biology: RNA discrimination. *Nature* **482,** 310–311. ISSN: 0028-0836 (2012).

Chapter 2

Non-coding yet non-trivial: a review on the computational genomics of lincRNAs

Travers Ching[1,2], Jason Masaki[3], Jason Weirather[4], Lana X. Garmire[1,2]

Published in *BioData Mining* (2015).

[1]University of Hawaii Cancer Center, 701 Ilalo Street, Honolulu, Hawaii, USA 96813

[2]Graduate Program of Molecular Biosciences and Bioengineering, University of Hawaii-Manoa, 1955 East-West Road, Honolulu, Hawaii, USA 96822

[3]Laboratory of Immunology and Signal Transduction, Chaminade University of Honolulu, Honolulu, HI 96816, USA

[4]Department of Internal Medicine, University of Iowa, Iowa City, IA, 52242, USA

2.1 Preface

Long intergenic non-coding RNAs (lincRNAs) represent one of the most mysterious RNA species encoded by the human genome. Thanks to next generation sequencing (NGS) technology and its applications, we have recently witnessed a surge in non-coding RNA research, including

lincRNA research. Here, we summarize the recent advancement in genomics studies of lincRNAs. We review the emerging characteristics of lincRNAs, the experimental and computational approaches to identify lincRNAs, their known mechanisms of regulation, the computational methods and resources for lincRNA functional predictions, and discuss the challenges to understanding lincRNA comprehensively.

2.2 Introduction

The mainstream focus of biomedical research has been in elucidating the functions and interactions among proteins within the cell. In line with the central dogma of molecular biology, RNAs were once perceived as the intermediary for protein production and the archaic precursor molecule of DNA. However, although RNAs are transcribed from more than 85% of genomic region [1], proteins are only encoded in less than 3% of human genome sequence [2]. This leaves a mysterious knowledge gap in either the efficiency of cellular transcription to translation or a foundational misunderstanding in gene expression regulation and RNA function. It was thought that RNAs had limited but essential and evolutionarily common roles of basic cell machinery such as tRNA, rRNA, and mRNA. The few examples of functional RNA or RNA with enzymatic-like activity, were considered as evolutionary remnant [3]. For a long period of time, non-coding RNA (ncRNA) transcripts were believed to be by-products derived from mRNA degradation or nonspecific polymerase activity, and therefore termed "transcriptional noise" [4].

It is now becoming evident that ncRNAs are responsible for many aspects of gene regulation. Some small non-coding RNAs, such as microRNAs, siRNAs, snRNAs, snoRNAs, exRNAs and piRNAs, have been well categorized over the past decade. However, long noncoding RNAs (lncRNAs) remained relatively unexplored due to the challenges of computational prediction under poor sequence conservation and low homology within the set of lncRNAs. However, some of these challenges have been addressed by the revolutionary inventions of next generation sequencing (NGS) and its applications, such as RNA-Seq, which capture whole transcriptome data, including lncRNAs. Among the human lncRNAs, tens of thousands of long intergenetic noncoding RNAs (lincRNAs) have been discovered in the genomic regions outside of the well-studied coding genomic regions, and they show many intriguing properties, such as associations with various human diseases, tissue-specific expression, and expression changes during development. Consequently, attributing organism complexity to the hidden regulation of lincRNAs is a fascinating new area of research. Here, we review the emerging characteristics of lincRNAs; the experimental and computational approaches to identifying lincRNAs and their mechanisms

of regulation; the challenges in computational predictions; and the resources still required to advance our understanding of lincRNA-related genetic regulation.

2.3 Review

2.3.1 Emerging characteristics of lincRNAs

LincRNAs are a putatively heterogeneous group, conventionally defined as ncRNA transcripts of more than 200 bp located in regions with no overlap to any known protein-coding genes. According to Lncipedia, a comprehensive lncRNA database, high-throughput studies of transcriptome data have catalogued over 111,000 lncRNA transcripts, with roughly 50% coming from intergenic region [5]. The majority of lincRNAs are thought to be transcribed from RNA polymerase II, and are therefore usually modified by post-transcriptional 5' capping and 3' polyadenylatio [6]. Surprisingly, lincRNAs show ribosome occupancy similar to the 5'UTRs of protein coding gene [7]. What differentiates lincRNAs from protein coding genes seems to be the lack of release upon encountering a stop codon in the lincRNA sequence [7]. Therefore, polyadenlyation and 5' capping are not necessarily markers of function. However, lincRNAs show a markedly higher degree of tissue-specific [8] and disease specific expressio [9], suggesting some biological function.

LincRNA expression is generally much lower than protein coding genes, with a few exceptions such as the XIST lincRN [10]. For some lincRNAs, even just a few or a single transcript exist in a cell, determined by RNA-Seq dat [10]. However, rather than being spurious by-products of non-specific RNA transcription, the expression levels of lincRNAs in any given cell are precisely coordinated throughout the tissue, and dynamic through the course of developmen [11]. Researchers have detected differential expression of lincRNA in a range of tissues, diseases, and specific cellular responses. Efforts have been made to take advantage of these properties of lincRNAs for translational and clinical applications, such as disease biomarker [12].

Another unique feature of lincRNAs is the low sequence conservation. LincRNAs exhibit 22-25% of conserved bases under purifying selection, compared to 77% in protein coding sequences. However, they are considerably more conserved than introns, which have 7% conservatio [13]. Under the assumption that sequence conservation reflects biological significance, the high genomic sequence variability in lincRNAs was the initial basis to call them "junk RNAs". Unlike proteins, where evolutionary conservation correlates highly with functional importance, lincRNAs seem to be under different selective pressures. Many lincRNA are predicted to have secondary structure and may therefore act in a sequence independent manne [14]. Consequently,

there may be a greater functional importance on molecular 3D conformation over the primary sequence. This is supported by a recent global study of genetic variants in human lincRNAs in association with diseases, where single nucleotide polymorphisms (SNPs) in evolutionarily conserved regions of lincRNAs had significant effects on predicted secondary structur [15].

2.3.2 Genome-wide detection of lincRNAs

Chromatin immunoprecipitation sequencing (ChIP-Seq) is an NGS method that has allowed the discovery of global genomic binding sites of DNA-interacting proteins, such as transcription factors and histones. Using ChIP-Seq signatures of histone 3 lysine 4 tri-methylation (H3K4me3) and histone 3 lysine 36 tri-methylation (HK36me3), or so called "K4-K36" clusters, Guttman et al. detected approximately 1700 transcriptional units $>$ 5kb among four mouse cell lines, which were confirmed by tiling microarrays, PCR and northern blot [16]. This type of chromatin signature was later applied to human cell lines to identify lincRNAs and was shown that along with HOTAIR, 20% of lincRNAs were associated with the Polycomb repressive complexes 2 (PRC2 [4]. ChIP-Seq has also been applied to the detection of RNA pol II occupancy to identify lincRNAs in mouse macrophages upon endotoxin stimulatio [17]. The authors found that 70% of extragenic polymerase II peaks were associated with genomic regions with a canonical chromatin signature of enhancers.

Clearly, decisions made during the library preparation phase of an RNA-seq experiment will affect lincRNA measurements. Since many but not all lincRNA transcripts are poly-adenylate [18], the decision to select poly-adenylated RNAs or to use ribodepletion methods should be made with care. Yang et al. [19] state that approximately 20% of transcripts are non-poly adenylated, suggesting that ribo-depletion methods are necessary to gain a more comprehensive picture of the transcriptome. In addition, Yang et al. find that some transcripts, such as the Malat1 lincRNA are bimorphic, meaning they exist in poly-A(+) and poly-A(-) configurations. Thus, ribo-depletion and poly-A selection methods could provide complementary information on the relative proportions of poly-adenylation of transcripts. Moreover, the adoption of strand-specific sequencing protocols provides a means of making more detailed annotations of lncRNAs, especially the antisense lncRNAs [20]. Nevertheless, even without strand information, RNA-seq has proven useful for the identification of lincRNAs. For example, Cabili et al. analysed lincRNAs in 24 tissues and mapped out nearly 9000 lincRNAs coupled to expression profile information [8].

Not all NGS methods are ideal for identifying the precise boundaries of lincRNAs. ChIP-Seq using antibodies against RNA polymerases can only provide a rough estimation of transcription

location but not the precise boundaries of transcripts [17]. RNA-Seq may also have trouble to detect isoforms and their exact start and end sites, as the cDNA is randomly fragmented, and accumulated from all isoforms within a given genomic loci [21]. Moreover, if RNA-Seq is conducted by a poly-A enriched approach, the internal bias against 5' ends make it difficult to map out the exact start sites of a transcript. However, some other NGS methods have been adopted to overcome this problem. For example, cap analysis gene expression (CAGE) tag sequencing has been used to aid the identification of transcription start sites in human cell [18], and 3'-end sequencing (3SEQ) has also been used in a zebrafish model to aid the determination of the 3' bounds of lincRNA transcript [22]. Additionally, tiling arrays that enable direct observation of lincRNAs transcript exons have been used to detect gene boundaries and alternative splicing. For example, Tahira et al. sampled intergenic and intronic ESTs from over one million ESTs from The Cancer Genome Project to develop a custom microarray, and subsequently identified lincRNAs differentially expressed between primary and metastatic pancreatic cancer [23].

2.3.3 Computational methods to predict lincRNAs

Most computational studies of lincRNAs rely on RNA-Seq results initially, with quality-control filtering steps to remove reads arising from spurious background nois [24]. Additional steps should be taken involving the removal of protein coding genes and small non-coding RNAs such as microRNAs. Methods to do such removals include ORF detection, BLAST to identify homologs of protein coding genes, domain based searches such as Pfa [1, 8, 9, 16], and predictions of coding potential based on nucleotide substitution frequencies given sequences from multiple species. The Coding Potential Calculator (CPC) [25] and iSeeRNA [26] programs are popular choices in determining coding potential. However, the extent to which some lincRNAs may be hosts of smaller RNA species such as microRNAs requires further study [27]. Another selection criterion is the number of exons in a transcript. Most of the exons (about 80% in human) are less than 200b [28], the minimum length requirement of lincRNA by definition. Transcripts with only one exon are less likely to be lincRNAs. Additionally, the number of exons can be used as an indicator of transcript quality. Multi-exonic transcripts are less likely to result from spurious transcription and genomic noise. The presence of introns is also indicative of robust and consistent transcription boundaries. Introns have less frequent terminal repeats and transposable elements in comparison to intergenic regions, suggesting that lincRNAs have additional conservation in splicing [29]. Finally, the axiomatic length-based filter, 200bp, eliminates any non-coding sequences that fall into the current small RNA categorie [30]. The filtering steps

described above are often implemented through a pipeline with a series of cut-offs or a decision tree to interrogate multiple features involved in classifying lincRNA [24].

In recent years, machine learning based classification approaches have been used to detect lincRNAs [17, 26, 31–33]. For example, iSeeRNA interrogated coding potential based on a variety of factors mentioned above, in addition to nucleotide composition. It was trained to differentiate protein coding genes and lncRNAs with an area under the curve (AUC) of 0.9 [26]. LncRNA-MFDL is another tool that uses a deep learning method and the fusion of multiple features to classify lincRNAs with an accuracy of 97.1 [32].

2.3.4 lincRNA databases

LincRNAs identified from exploratory studies are a valuable resource for accumulating information about these relatively unknown transcripts. Information such as location, splice junction, and tissue specificity are important features. There are quite a few specialized databases that provide comprehensive annotations for lincRNAs or lncRNAs. These include The Broad institute's Human Body Map projec [8], NONCODE [34] and Lncipedi [5]. Other large gene annotation sets such as GENCODE [35, 36], UCSC's known genes [37] or Rfam [38] RNA family databases are not specific to non-coding RNAs, but nevertheless contain large sets of annotations and information on lincRNAs.

The UCSC ENCODE project provides a feature-rich resource to describe the transcriptional landscape in a variety of tissues from the GENCODE databas [39]. The Ensembl Geneome Browser is another resource that identifies and annotates transcripts within their large database using transcriptional evidence as well as chromatin mark-up [35]. The Ensembl project uses the GENCODE database, and contributes multiple sources to GENCODE through an automated annotation pipeline in combination with the large Havana annotation by the Sanger Institut [35]. While GENCODE is one of the most comprehensive databases for mammalian species, it does not include lincRNAs found by RNA-Seq ab initio alignment methods, such as those in the Human Body Map. Neither is it as comprehensive as specialized databases.

More specialized lncRNA databases, such as NONCODE and Lncipedia, enumerate a much larger number of lncRNAs (Table 1). These databases have been created to facilitate functional analyses by integrating multiple data sources such as expression, chromatin markups, microRNA binding sites and mutational data with known lncRNAs. Not surprisingly, the overlap of those data sets can differ greatly, largely due to the selection criteria of particular lncRNAs or the tissue origins where lincRNAs were initially detected.

2.3.5 Genomic assays to study lincRNA regulations

2.3.5.1 Methods to elucidate the functions

of individual lincRNAs have made much slower progress compared to large-scale genomic assays. In this section we survey the increasing number of genome-scale molecular interaction studies to investigate the cellular functions of lincRNAs. Several genomic approaches have been reported to identify specific functions of lincRNAs. One popular technique is the protein-centric RNA immunoprecipitation (RIP), which selects a particular protein or a group of proteins to co-precipitate RNAs and determines functional relationships based on physical interaction [40]. This allows one to ascribe functions of the protein(s) with co-precipitated lincRNAs. For example, Shi et al. used RIP to identify novel functional lincRNAs involved in the regulation of TNF expression through binding to PRC [41], and found that PRC2 binds to thousands of RNA species. Thus, protein-centric methods focusing on PRC2 have provided us critical insights into the genome-wide regulation by lincRNA [42].

Conversely, another approach is to purify certain RNA molecules and then capture the associated proteins (RNA-centric methods); the associated proteins can then be identified via mass-spectroscop [40]. This approach works by complementary base pairing of the RNA sequence to oligonucleotide probes labelled with streptavidin or bioti [43]. However, in comparison to protein-centric methods where the RNA targets can be amplified by PCR, RNA-centric methods do not have a means of amplifying the protein targets. Therefore, RNA-centric methods work best when large quantities of protein are availabl [40].

Additionally, there have also been a handful of "DNA-centric" methods for studying lincRNAs. Methods that investigate DNA modification or the 3D structure of chromosomes have greatly advanced our understanding of gene regulatio [44]. For example, Ma et al. developed a novel method called Dnase Hi-C that determines the interactions of lincRNA promoters with DNA enhancer region [44]. Their method involves cross-linking nearby DNA strands, followed by DNAse I digestion, proximity ligation between the cross-linked strands and DNA sequencing. Rather than using restrictive enzyme (RE) as done in conventional Hi-C, which generates predictable and consistent fragment ends, DNase I produces a heterogeneous mixture of fragment ends that greatly improves the efficiency and resolution. They were able to fine-map cell specific 3D organization of 998 lincRNA promoters. They demonstrated that lincRNA expression is tightly controlled by complex mechanisms including super-enhancers and PRCs. Known functions and mechanisms of lincRNAs

Historically, lincRNAs have been shown to have a greater likelihood to be functionally associated with their nearest neighbouring protein-coding genes. However, more recent analyses show that the expression correlation between a lincRNA and its closest coding gene is not statistically significant when compared to the correlation between two neighbouring protein-coding gene [8, 45]. While complementary base pairing may be the mechanism of action for some small RNAs such as microRNAs, lincRNAs by their nature are unlikely to exert their regulatory function solely through sequence pairing. Instead, lincRNAs have been shown to mediate the interplay between many molecular species simultaneousl [46]. LincRNAs affect gene expression by many different mechanisms – from chromatin remodelling and epigenetic regulation, to transcriptional, post-transcriptional, and protein-level control. So far, no unifying genome-wide theme has been found to explain all the complexities of lincRNA regulation. We review the handful of competing theories that attempt to address this problem.

a. LincRNAs involved in chromatin remodelling

Epigenetics is a vital means of DNA patterning to regulate gene expressio [47]. PRCs exert gene silencing epigenetically by histone modifications and DNA chemical alterations such as methylation [42]. Recruitment of PRCs to certain genomic locations is mediated by specific lincRNAs. Thus, the differential expression of certain lincRNAs (such as HOTAIR) can lead to activation or deactivation of transcription on the genom [48]. The vital role of gene suppression due to lincRNAs has been implicated in the pathology of cancers, where dysregulation of individual lincRNAs release cell cycle control resulting in an increase in cell proliferatio [49]. Complicating matters, thousands of lincRNAs were found bound by PRC2 within various cell types [4], suggesting the widespread interaction of lincRNAs with the epigenetic modification machinery.

b. LincRNAs as transcription co-factors

Many lincRNAs are known to act as transcription co-factors. In some cases, the act of transcription of a lincRNA may positively or negatively affect expression of nearby genes [50]. Dimitrova et al. showed that lincRNA-p21 acts as a transcriptional coactivator and was required for recruitment of ribonucleoproteins to promoter elements associated with pre-mRN [51]. MALAT1 is also known to act as a transcription co-factor. This lincRNA is well characterized as one of the most highly expressed mammalian lincRNAs. It is also known to significantly affect the metastasic process in lung adenocarcinoma, by enhancing the expression of cell motility genes [52]. It was found that MALAT1 acts as a molecular scaffold to allow gene expression

by promoting the interaction among unmethylated PRC2, E2F1 transcription factor, histone markers, and the other transcriptional co-activator complexe [53]. Interestingly, this protein sequestration mechanism of ncRNA is not unique to eukaryotes, and it also occurs in bacteri [54].

c. Competing endogenous RNA hypothesis of lincRNAs

The competing endogenous RNA (ceRNA) hypothesis is a theory that lncRNAs (including lincRNAs) regulate gene expression by acting as microRNA sponge [55]. The inhibition of specific mRNA translation is modulated by microRNA depletion through lncRNAs harbouring microRNA binding sites. By effectively competing for the same microRNA, these lncRNAs exert a level of competitive inhibition. Based on this hypothesis, Liu et al developed a database of lincRNAs that were predicted to have functional associations with protein-coding genes [56]. Some exemplary lincRNAs that function as ceRNAs are the HUL [57] and LINC-RO [33]. HULC was shown to be the molecular sponge of a series of microRNAs including miR-372, which induces phosphorylation of CREB in liver cancer [57], and LINC-ROR shares the microRNA response elements with core transcription factors Oct4, Sox2, and Nanog and thus increases expression of these genes by competing for microRNAs [58]. Although some lincRNAs act as ceRNAs, it is unclear how prevalent this mechanism is among all lincRNAs.

d. LincRNAs as evolutionary reservoirs

While lincRNAs have less sequence conservation than protein-coding genes, they have a greater degree of secondary motif conservation compared to mRNA [59]. These elements may explain the origins of lincRNAs, which provide a reservoir of evolutionarily constrained RNA motif [59, 60] to supply extra genetic modules for evolutionary tinkering. It is also known that Retrotransposon and tandem repeat sequences are more common within lincRNAs compared to protein-coding gene [61]. Embedded microRNAs and the hypothesized ceRNA mechanism mentioned earlier may be accounted for by such duplication events, as modulating copy number of an embedded microRNA or target site would allow for fine-tuned regulatio [55, 62].

Computational methods for lincRNA target prediction

There have been many attempts to computationally identify the function of lincRNAs. Given the length of lincRNA sequences and the complexity of their potential 3D structures along

with the RNA and protein partners, this is a very challenging task. We review the different computational approaches in the following.

a. Correlation with protein coding genes and biological processes

One of the simplest approaches to determine the function of lincRNAs is to examine their correlations with protein coding genes [63]. However, this is a "black box" approach that identifies neither causality nor lincRNA functions at the molecular level. Another naive approach is to relate the function of lincRNAs to the nearby protein coding gene [64]. Many lincRNAs have been found to exert regulatory activity on protein coding genes in cis [44, 51]. However, Khalil et al. found that knockdown of six different lincRNAs did not affect the expression of level of nearby gene [4]. This suggests that lincRNAs can work in trans as well, and that the correlation between a lincRNA and its nearby protein coding genes may not necessarily be a causative relationship, but rather a result of sharing a region of active transcription.

b. Relation between lincRNAs with microRNAs and other small non-coding RNAs

Other more sophisticated tools have been developed to identify more succinct functions. Boerner and McGinnis constructed a pipeline to seek functions of lncRNAs in Zea May [31]. Using BLAST search, they found that the majority of lncRNAs have strong homology to small RNA molecules. They hypothesized that many lncRNAs are simply unprocessed pre-cursors to small non-coding RNAs, such as microRNA, shRNA and siRN [31]. Based on the "ceRNA hypothesis" mentioned earlier, Liu et al developed "linc2GO", a software for identifying mRNA and lincRNA pair [56]. Using predicted microRNA targets from miRanda, TargetScan and PITA software, they predicted microRNA targets on both mRNAs and lincRNAs; The mRNAs and lincRNAs that had statistically significant target sites for a particular microRNA were proposed to have a "competing endogenous" relationship.

c. Machine learning approaches to target and functional prediction

Machine learning methods have been used successfully to classify whether transcripts are coding or non-coding. However, machine learning methods to identify the targets of lincRNAs have not seen much success. Comparatively, there has been much more success in using supervised learning approaches to identify microRNA targets, such as TargetSca [65], SvMicr [66] and mirMar [67]. Still progress is being made towards lincRNA functional prediction. Glazko et

al. used support vector machines (SVM) to predict lncRNA and PRC2 binding using human lincRNA associated with PRC2 as training data. With the classification model, they were able to predict 59.4% of lncRNAs which bind to PRC2 in mice [68]. The model was based off of the dataset by Khalil et al. [4] which found roughly 20% of lincRNAs to associate with PRC2. However, it remains unclear whether the associations were spurious or led to sequence specific chromatin regulation.

d. LincRNA functional prediction through the higher-order structure

Perhaps the least explored lincRNA prediction approach is functional prediction through tertiary and quaternary structure. As the structure of RNA molecules are related to their functions, predicting the structure of complexes between RNA-RNA, and RNA-protein interactions could elucidate functional properties. Several RNA-RNA interaction prediction tools are available, usually based on free-energy, such as RNAhybri [69] and RNADuple [**SEBASTIANFast**]. RNA-protein interaction prediction tools exist as well, such as RPIseq which uses a Random Forest classification approac [70] or RNApred, which uses an SVM approac [71]. However, there have not been many attempts for lincRNA functional prediction. Many of the protein complexes interacting with lincRNAs do not fall into common binding motif [40]. Furthermore, functional prediction is complicated by the "n-body problem", since protein, RNA and DNA can be complexed with lincRNAs simultaneously.

e. Downstream target prediction through directed graphs

Reverse engineering of gene regulatory networks has been an area of research before the explosion of next generation sequencing and lincRNA researc [72]. Approaches such as Bayesian networks, information-theoretic approaches and ordinary differential equations have shown strong performanc [73]. Generally, a perturbation of the system (such as gene knockout, overexpression or drug treatment) is performed which forces a node (i.e., a gene) on a regulatory network graph to be forcibly turned on or turned off. This perturbation produces direct causative (rather than correlative) downstream effects that can be captured through microarrays and quantitative methods. Recently, Jiang et al. published a database (lncRNA2Target) describing lncRNA knockdown and overexpression experiments, followed by gene quantification by microarray or qPC [74]. These types of experiments can be a valuable resource for elucidating a lincRNA's targets and pathways.

2.4 Conclusion

Statistical evaluation studies for lincRNAs are urgently needed, as datasets produced by these various methods have thus far shown only modest overlaps in their identified lincRNAs [14] Besides lack of sequence conservation among lincRNAs, another major issue hindering functional prediction is the lack of validated data. While there are many well-studied lincRNAs, there are massively more unannotated lincRNAs. Machine learning methods often require a large training dataset to produce accurate results. Several functional lincRNA/lncRNA databases exist (such as lncrnaDB), however the number of entries are very low and do not categorize the function of the lncRNAs in a systematic manne [75]. As more and more lincRNAs become functionally validated, comprehensive and regularly updated databases would be a great source to build good prediction methods. Perhaps even more important is the advancement of experimental techniques to provide quality data required for the prediction. Currently, most experimental techniques focus on a single protein or a small number of proteins (protein-centric) or a single lincRNA or family of lincRNAs (RNA-centric [40]. New methods are required that can provide high-throughput protein and RNA targets of thousands of lincRNAs in parallel.

2.5 Acknowledgements

This work was supported by grants K01ES025434 awarded by NIEHS through funds provided by the trans-NIH Big Data to Knowledge (BD2K) initiative (www.bd2k.nih.gov), P20 COBRE GM103457 awarded by NIH/NIGMS, and Medical Research Grant 14ADVC-64566 from Hawaii Community Foundation to L.X. Garmire.

2.6 Competing interests

The authors declare that they have no competing interests.

2.7 Authors' contributions

LXG planned the work. TC, JM, JW and LXG all wrote parts of the manuscript. TC and LXG designed and finalized the manuscript.

References

1. Hangauer, M. J., Vaughn, I. W. & McManus, M. T. Pervasive transcription of the human genome produces thousands of previously unidentified long intergenic noncoding RNAs. *PLoS genetics* **9,** e1003569. ISSN: 1553-7404 (2013).
2. Ezkurdia, I., Juan, D., Rodriguez, J. M., Frankish, A., Diekhans, M., Harrow, J., Vazquez, J., Valencia, A. & Tress, M. L. Multiple evidence strands suggest that there may be as few as 19 000 human protein-coding genes. *Hum Mol Genet.* ISSN: 1460-2083 (Electronic) 0964-6906 (Linking). doi:10.1093/hmg/ddu309 (2014).
3. Joyce, G. F. The antiquity of RNA-based evolution. *Nature* **418,** 214–221. ISSN: 0028-0836 (2002).
4. Khalil, A. M., Guttman, M., Huarte, M., Garber, M., Raj, A., Morales, D. R., Thomas, K., Presser, A., Bernstein, B. E. & van Oudenaarden, A. Many human large intergenic noncoding RNAs associate with chromatin-modifying complexes and affect gene expression. *Proceedings of the National Academy of Sciences* **106,** 11667–11672. ISSN: 0027-8424 (2009).
5. Volders, P.-J., Verheggen, K., Menschaert, G., Vandepoele, K., Martens, L., Vandesompele, J. & Mestdagh, P. An update on LNCipedia: a database for annotated human lncRNA sequences. *Nucleic acids research* **43,** D174–D180. ISSN: 0305-1048 (2015).
6. Ulitsky, I., Shkumatava, A., Jan, C. H., Sive, H. & Bartel, D. P. Conserved function of lincRNAs in vertebrate embryonic development despite rapid sequence evolution. *Cell* **147,** 1537–50. ISSN: 1097-4172 (Electronic) 0092-8674 (Linking) (2011).
7. Guttman, M., Russell, P., Ingolia, N. T., Weissman, J. S. & Lander, E. S. Ribosome profiling provides evidence that large noncoding RNAs do not encode proteins. *Cell* **154,** 240–251. ISSN: 0092-8674 (2013).
8. Cabili, M. N., Trapnell, C., Goff, L., Koziol, M., Tazon-Vega, B., Regev, A. & Rinn, J. L. Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. *Genes Dev* **25,** 1915–27. ISSN: 1549-5477 (Electronic) 0890-9369 (Linking) (2011).
9. Iyer, M. K., Niknafs, Y. S., Malik, R., Singhal, U., Sahu, A., Hosono, Y., Barrette, T. R., Prensner, J. R., Evans, J. R. & Zhao, S. The landscape of long noncoding RNAs in the human transcriptome. *Nature Genetics.* ISSN: 1061-4036 (2015).

10. Mercer, T. R., Gerhardt, D. J., Dinger, M. E., Crawford, J., Trapnell, C., Jeddeloh, J. A., Mattick, J. S. & Rinn, J. L. Targeted RNA sequencing reveals the deep complexity of the human transcriptome. *Nat Biotechnol* **30,** 99–104. ISSN: 1546-1696 (Electronic) 1087-0156 (Linking) (2012).

11. Sauvageau, M., Goff, L. A., Lodato, S., Bonev, B., Groff, A. F., Gerhardinger, C., Sanchez-Gomez, D. B., Hacisuleyman, E., Li, E. & Spence, M. Multiple knockout mouse models reveal lincRNAs are required for life and brain development. *Elife* **2,** e01749. ISSN: 2050-084X (2013).

12. Ge, X., Chen, Y., Liao, X., Liu, D., Li, F., Ruan, H. & Jia, W. Overexpression of long noncoding RNA PCAT-1 is a novel biomarker of poor prognosis in patients with colorectal cancer. *Medical oncology* **30,** 1–6. ISSN: 1357-0560 (2013).

13. Guttman, M. *et al.* Ab initio reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs. *Nature Biotechnology* **28,** 503–U166. ISSN: 1087-0156 (2010).

14. Marques, A. C. & Ponting, C. P. Catalogues of mammalian long noncoding RNAs: modest conservation and incompleteness. *Genome Biol* **10,** R124 (2009).

15. Ning, S., Wang, P., Ye, J., Li, X., Li, R., Zhao, Z., Huo, X., Wang, L., Li, F. & Li, X. A global map for dissecting phenotypic variants in human lincRNAs. *European Journal of Human Genetics* **21,** 1128–1133. ISSN: 1018-4813 (2013).

16. Guttman, M. *et al.* Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. *Nature* **458,** 223–7. ISSN: 1476-4687 (Electronic) 0028-0836 (Linking) (2009).

17. Garmire, L. X., Garmire, D. G., Huang, W., Yao, J., Glass, C. K. & Subramaniam, S. A global clustering algorithm to identify long intergenic non-coding RNA–with applications in mouse macrophages. *PLoS One* **6,** e24051. ISSN: 1932-6203 (Electronic) 1932-6203 (Linking) (2011).

18. Djebali, S. *et al.* Landscape of transcription in human cells. *Nature* **489,** 101–8. ISSN: 1476-4687 (Electronic) 0028-0836 (Linking) (2012).

19. Yang, L., Lin, C., Liu, W., Zhang, J., Ohgi, K. A., Grinstein, J. D., Dorrestein, P. C. & Rosenfeld, M. G. ncRNA- and Pc2 methylation-dependent gene relocation between nuclear structures mediates gene activation programs. *Cell* **147,** 773–88. ISSN: 1097-4172 (Electronic) 0092-8674 (Linking) (2011).

20. He, Y., Vogelstein, B., Velculescu, V. E., Papadopoulos, N. & Kinzler, K. W. The antisense transcriptomes of human cells. *Science* **322,** 1855–7. ISSN: 1095-9203 (Electronic) 0036-8075 (Linking) (2008).

21. Kawaji, H., Lizio, M., Itoh, M., Kanamori-Katayama, M., Kaiho, A., Nishiyori-Sueki, H., Shin, J. W., Kojima-Ishiyama, M., Kawano, M. & Murata, M. Comparison of CAGE and RNA-seq transcriptome profiling using clonally amplified and single-molecule next-generation sequencing. *Genome research* **24,** 708–717. ISSN: 1088-9051 (2014).

22. Ulitsky, I. & Bartel, D. P. lincRNAs: genomics, evolution, and mechanisms. *Cell* **154,** 26–46. ISSN: 0092-8674 (2013).

23. Tahira, A. C., Kubrusly, M. S., Faria, M. F., Dazzani, B., Fonseca, R. S., Maracaja-Coutinho, V., Verjovski-Almeida, S., Machado, M. C. & Reis, E. M. Long noncoding intronic RNAs are differentially expressed in primary and metastatic pancreatic cancer. *Mol Cancer* **10,** 141. ISSN: 1476-4598 (Electronic) 1476-4598 (Linking) (2011).

24. Prensner, J. R., Iyer, M. K., Balbin, O. A., Dhanasekaran, S. M., Cao, Q., Brenner, J. C., Laxman, B., Asangani, I. A., Grasso, C. S. & Kominsky, H. D. Transcriptome sequencing across a prostate cancer cohort identifies PCAT-1, an unannotated lincRNA implicated in disease progression. *Nature biotechnology* **29,** 742–749. ISSN: 1087-0156 (2011).

25. Kong, L., Zhang, Y., Ye, Z. Q., Liu, X. Q., Zhao, S. Q., Wei, L. & Gao, G. CPC: assess the protein-coding potential of transcripts using sequence features and support vector machine. *Nucleic Acids Res* **35,** W345–9. ISSN: 1362-4962 (Electronic) 0305-1048 (Linking) (2007).

26. Sun, K., Chen, X., Jiang, P., Song, X., Wang, H. & Sun, H. iSeeRNA: identification of long intergenic non-coding RNA transcripts from transcriptome sequencing data. *BMC genomics* **14,** S7. ISSN: 1471-2164 (2013).

27. Jalali, S., Jayaraj, G. G. & Scaria, V. Integrative transcriptome analysis suggest processing of a subset of long non-coding RNAs to small RNAs. *Biol Direct* **7,** 25. ISSN: 1745-6150 (Electronic) 1745-6150 (Linking) (2012).

28. Sakharkar, M. K., Chow, V. T. & Kangueane, P. Distributions of exons and introns in the human genome. *In Silico Biol* **4,** 387–93. ISSN: 1386-6338 (Print) 1386-6338 (Linking) (2004).

29. Semon, M. & Duret, L. Evidence that functional transcription units cover at least half of the human genome. *Trends in Genetics* **20,** 229–232. ISSN: 0168-9525 (2004).

30. Qiu, M. T., Hu, J. W., Yin, R. & Xu, L. Long noncoding RNA: an emerging paradigm of cancer research. *Tumour Biol* **34,** 613–20. ISSN: 1423-0380 (Electronic) 1010-4283 (Linking) (2013).

31. Boerner, S. & McGinnis, K. M. Computational identification and functional predictions of long noncoding RNA in Zea mays. *PloS one* **7,** e43047. ISSN: 1932-6203 (2012).

32. Fan, X.-N. & Zhang, S.-W. lncRNA-MFDL: identification of human long non-coding RNAs by fusing multiple features and using deep learning. *Molecular BioSystems* (2015).

33. Wang, Y., Xu, Z., Jiang, J., Xu, C., Kang, J., Xiao, L., Wu, M., Xiong, J., Guo, X. & Liu, H. Endogenous miRNA sponge lincRNA-RoR regulates Oct4, Nanog, and Sox2 in human embryonic stem cell self-renewal. *Dev Cell* **25,** 69–80. ISSN: 1878-1551 (Electronic) 1534-5807 (Linking) (2013).

34. Xie, C., Yuan, J., Li, H., Li, M., Zhao, G., Bu, D., Zhu, W., Wu, W., Chen, R. & Zhao, Y. NONCODEv4: exploring the world of long non-coding RNA genes. *Nucleic acids research* **42,** D98–D103. ISSN: 0305-1048 (2014).

35. Flicek, P. *et al.* Ensembl 2012. *Nucleic Acids Res* **40,** D84–90. ISSN: 1362-4962 (Electronic) 0305-1048 (Linking) (2012).

36. Harrow, J. *et al.* GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Res* **22,** 1760–74. ISSN: 1549-5469 (Electronic) 1088-9051 (Linking) (2012).

37. Hsu, F., Kent, W. J., Clawson, H., Kuhn, R. M., Diekhans, M. & Haussler, D. The UCSC known genes. *Bioinformatics* **22,** 1036–1046. ISSN: 1367-4803 (2006).

38. Griffiths-Jones, S., Moxon, S., Marshall, M., Khanna, A., Eddy, S. R. & Bateman, A. Rfam: annotating non-coding RNAs in complete genomes. *Nucleic acids research* **33,** D121–D124. ISSN: 0305-1048 (2005).

39. Dunham, I. *et al.* An integrated encyclopedia of DNA elements in the human genome. *Nature* **489,** 57–74. ISSN: 1476-4687 (Electronic) 0028-0836 (Linking) (2012).

40. McHugh, C. A., Russell, P. & Guttman, M. Methods for comprehensive experimental identification of RNAa-protein interactions. *Genome Biol* **15,** 203 (2014).

41. Shi, L., Song, L., Fitzgerald, M., Maurer, K., Bagashev, A. & Sullivan, K. E. Noncoding RNAs and LRRFIP1 regulate TNF expression. *J Immunol* **192,** 3057–67. ISSN: 1550-6606 (Electronic) 0022-1767 (Linking) (2014).

42. Goff, L. A. & Rinn, J. L. Poly-combing the genome for RNA. *Nature structural & molecular biology* **20,** 1344–1346. ISSN: 1545-9993 (2013).

43. Gong, C. & Maquat, L. E. *Affinity Purification of Long Noncoding RNA-Protein Complexes from Formaldehyde Cross-Linked Mammalian Cells* 81–86. ISBN: 1493913689 (Springer, 2015).

44. Ma, W., Ay, F., Lee, C., Gulsoy, G., Deng, X., Cook, S., Hesson, J., Cavanaugh, C., Ware, C. B. & Krumm, A. Fine-scale chromatin interaction maps reveal the cis-regulatory landscape of human lincRNA genes. *Nature methods.* ISSN: 1548-7091 (2014).

45. Wright, M. W. A short guide to long non-coding RNA gene nomenclature. *Human Genomics* **8.** ISSN: 1473-9542. doi:Doi10.1186/1479-7364-8-7 (2014).

46. Mercer, T. R., Dinger, M. E. & Mattick, J. S. Long non-coding RNAs: insights into functions. *Nature Reviews Genetics* **10,** 155–159. ISSN: 1471-0056 (2009).

47. Di Croce, L. & Helin, K. Transcriptional regulation by Polycomb group proteins. *Nat Struct Mol Biol* **20,** 1147–55. ISSN: 1545-9985 (Electronic) 1545-9985 (Linking) (2013).

48. Loewen, G., Zhuo, Y., Zhuang, Y., Jayawickramarajah, J. & Shan, B. lincRNA HOTAIR as a novel promoter of cancer progression. *Journal of Cancer Research Updates* **3,** 134–140. ISSN: 1929-2279 (2014).

49. Gutschner, T. & Diederichs, S. The hallmarks of cancer: a long non-coding RNA point of view. *RNA Biol* **9,** 703–19. ISSN: 1555-8584 (Electronic) 1547-6286 (Linking) (2012).

50. Wilusz, J. E., Sunwoo, H. & Spector, D. L. Long noncoding RNAs: functional surprises from the RNA world. *Genes Dev* **23,** 1494–504. ISSN: 1549-5477 (Electronic) 0890-9369 (Linking) (2009).

51. Dimitrova, N., Zamudio, J. R., Jong, R. M., Soukup, D., Resnick, R., Sarma, K., Ward, A. J., Raj, A., Lee, J. T., Sharp, P. A. & Jacks, T. LincRNA-p21 activates p21 in cis to promote Polycomb target gene expression and to enforce the G1/S checkpoint. *Mol Cell* **54,** 777–90. ISSN: 1097-4164 (Electronic) 1097-2765 (Linking) (2014).

52. Wang, K. C. & Chang, H. Y. Molecular Mechanisms of Long Noncoding RNAs. *Molecular Cell* **43,** 904–914. ISSN: 1097-2765 (2011).

53. Yang, L., Duff, M. O., Graveley, B. R., Carmichael, G. G. & Chen, L.-L. Genomewide characterization of non-polyadenylated RNAs. *Genome Biol* **12,** R16 (2011).

54. Duss, O., Michel, E., Yulikov, M., Schubert, M., Jeschke, G. & Allain, F. H. T. Structural basis of the non-coding RNA RsmZ acting as a protein sponge. *Nature* **509,** 588–+. ISSN: 0028-0836 (2014).

55. Salmena, L., Poliseno, L., Tay, Y., Kats, L. & Pandolfi, P. P. A ceRNA hypothesis: the Rosetta Stone of a hidden RNA language? *Cell* **146,** 353–8. ISSN: 1097-4172 (Electronic) 0092-8674 (Linking) (2011).

56. Liu, K., Yan, Z., Li, Y. & Sun, Z. Linc2GO: a human LincRNA function annotation resource based on ceRNA hypothesis. *Bioinformatics* **29,** 2221–2222. ISSN: 1367-4803 (2013).

57. Wang, J., Liu, X., Wu, H., Ni, P., Gu, Z., Qiao, Y., Chen, N., Sun, F. & Fan, Q. CREB up-regulates long non-coding RNA, HULC expression through interaction with microRNA-372 in liver cancer. *Nucleic Acids Res* **38,** 5366–83. ISSN: 1362-4962 (Electronic) 0305-1048 (Linking) (2010).

58. Wang, Y., Li, Y., Wang, Q., Lv, Y., Wang, S., Chen, X., Yu, X., Jiang, W. & Li, X. Computational identification of human long intergenic non-coding RNAs using a GA-SVM algorithm. *Gene* **533,** 94–99. ISSN: 0378-1119 (2014).

59. Kapusta, A. & Feschotte, C. Volatile evolution of long noncoding RNA repertoires: mechanisms and biological implications. *Trends in Genetics* **30,** 439–452. ISSN: 0168-9525 (2014).

60. Smith, M. A., Gesell, T., Stadler, P. F. & Mattick, J. S. Widespread purifying selection on RNA structure in mammals. *Nucleic acids research* **41,** 8220–8236. ISSN: 0305-1048 (2013).

61. Kelley, D. & Rinn, J. Transposable elements reveal a stem cell-specific class of long noncoding RNAs. *Genome Biology* **13.** ISSN: 1474-7596. doi:Doi10.1186/Gb-2012-13-11-R107 (2012).

62. Labialle, S. & Cavaille, J. Do repeated arrays of regulatory small-RNA genes elicit genomic imprinting? *Bioessays* **33,** 565–573. ISSN: 1521-1878 (2011).

63. Guo, X., Gao, L., Liao, Q., Xiao, H., Ma, X., Yang, X., Luo, H., Zhao, G., Bu, D. & Jiao, F. Long non-coding RNAs function annotation: a global prediction method based on bi-colored networks. *Nucleic acids research* **41,** e35–e35. ISSN: 0305-1048 (2013).

64. Ma, H., Hao, Y., Dong, X., Gong, Q., Chen, J., Zhang, J. & Tian, W. Molecular mechanisms and function prediction of long noncoding RNA. *The Scientific World Journal* **2012** (2012).

65. Lewis, B. P., Burge, C. B. & Bartel, D. P. Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. *cell* **120,** 15–20. ISSN: 0092-8674 (2005).

66. Liu, H., Yue, D., Chen, Y., Gao, S.-J. & Huang, Y. Improving performance of mammalian microRNA target prediction. *BMC bioinformatics* **11,** 476. ISSN: 1471-2105 (2010).

67. Menor, M., Ching, T., Zhu, X., Garmire, D. & Garmire, L. X. mirMark: a site-level and UTR-level classifier for miRNA target prediction. *Genome biology* **15,** 500. ISSN: 1465-6906 (2014).

68. Glazko, G. V., Zybailov, B. L. & Rogozin, I. B. Computational prediction of polycomb-associated long non-coding RNAs. *PloS one* **7,** e44878. ISSN: 1932-6203 (2012).

69. Kruger, J. & Rehmsmeier, M. RNAhybrid: microRNA target prediction easy, fast and flexible. *Nucleic acids research* **34,** W451–W454. ISSN: 0305-1048 (2006).

70. Muppirala, U., Lewis, B. A. & Dobbs, D. Computational tools for investigating RNA-protein interaction partners. *J Comput Sci Syst Biol* **6,** 182–187 (2013).

71. Kumar, M., Gromiha, M. M. & Raghava, G. P. SVM based prediction of RNA-binding proteins using binding residues and evolutionary information. *Journal of Molecular Recognition* **24,** 303–313. ISSN: 1099-1352 (2011).

72. Murphy, K. & Mian, S. *Modelling gene expression data using dynamic Bayesian networks* Report (Technical report, Computer Science Division, University of California, Berkeley, CA, 1999).

73. Bansal, M., Belcastro, V., Ambesi-Impiombato, A. & Di Bernardo, D. How to infer gene networks from expression profiles. *Molecular systems biology* **3.** ISSN: 1744-4292 (2007).

74. Jiang, Q., Wang, J., Wu, X., Ma, R., Zhang, T., Jin, S., Han, Z., Tan, R., Peng, J. & Liu, G. LncRNA2Target: a database for differentially expressed genes after lncRNA knockdown or overexpression. *Nucleic acids research* **43,** D193–D196. ISSN: 0305-1048 (2015).

75. Galperin, M. Y., Rigden, D. J. & Fernandez-Suarez, X. M. The 2015 Nucleic Acids Research Database Issue and Molecular Biology Database Collection. *Nucleic acids research* **43,** D1–D5. ISSN: 0305-1048 (2015).

Table 1: Summary of lncRNA/lincRNA databases

Project Name	Species	Purpose
Human Body Map	Human	A reference set of lincRNAs
ChIPBase	Various (incl. Human and Mouse)	A resource for lncRNA transcriptional regulation and expression profiles of ncRNA (lncRNA, microRNAs, etc.)
NONCODE	Various (incl. Human and Mouse)	A large lncRNA database integrating various databases and references
lncRNAdb	Various (incl. Human and Mouse)	A database of lncRNAs having biological function or regulatory roles
ncRNA expression database (NRED)	Human and Mouse	Expression database for human and mouse lncRNAs
LNCipedia	Various (incl. Human and Mouse)	A large database of lncRNA transcripts and annotation
LncRNADisease	Human	A database of lncRNAs associated with human diseases
DIANA-LncBase	Human and Mouse	A database of experimentally verified and predicted microRNA targets on lncRNAs
lncRNA2Target	Human and Mouse	A collection of lncRNA knockout experiments and downstream regulation
starBase 2.0	Human, Mouse and C. elegans	A collection of lncRNA and predicted microRNA targets; lncRNA expression profiles from TCGA data
lncRNAMap	Human	A resource for exploring lncRNA expression profiles and interaction with small RNAs (siRNA, microRNAs, etc.)
lncRNAWiki	Human	An open wiki style lncRNA database
MONOCLdb	Mouse	A mouse noncoding database detailing functional enrichment of lncRNA in response to respiratory disease caused by influenza and SARS-CoV
lncRNome	Human	A searchable database for long noncoding RNAs in humans and various properties, such as predicted structure, SNPs and epigenetic modifications
PLncDB	Arabidopsis thaliana	A database dedicated to A. thaliana plant lncRNA transcriptome, including information on epigenetic modification
Functional lncRNA Database	Human, Mouse and Rat	A database of experimentally validated functional lncRNAs
lnCeDB	Human	A database of lncRNA acting as ceRNA
Linc2GO	Human	A database of lncRNA acting as ceRNA and biological processes based on GO annotation
lncRNASNP	Human and Mouse	A database cataloging micro-RNA interactions and SNPs in lncRNAs and their impact on secondary structure

2.8 Chapter summary

In this chapter, we review the current state of lincRNA research from the perspective of computational genomics. How has high throughput sequencing and subsequent computational analysis changed the study of non-coding RNAs? Through sequencing methods such as CLIP-Seq and RNA-Seq, thousands of novel lincRNAs have been identified in recent years. The goal of this dissertation is the application of these methods to investigate lincRNA relation to cancer research.

Thus, better understanding the field has helped accomplish the research in later chapters. For example, in chapter 4 and chapter 7, we used genomic annotation derived from early lincRNA studies in order to quantify lincRNAs in RNA-Seq samples. Through extensive literature review, we find that there remains many additional challenges to lincRNA genomic research. Such challenges include better understanding the both the breadth and depth of the lincRNA molecular mechanisms, better predicting their roles in biological pathways and building better databases related to lincRNA function and metadata.

Chapter 3

Power Analysis and Sample Size Estimation for RNA-Seq Differential Expression

Travers Ching[1,2], Sijia Huang[1,2], Lana X. Garmire[1,2]

Published in *RNA* (2014).

[1]University of Hawaii Cancer Center, 701 Ilalo Street, Honolulu, Hawaii, USA 96813

[2]Graduate Program of Molecular Biosciences and Bioengineering, University of Hawaii-Manoa, 1955 East-West Road, Honolulu, Hawaii, USA 96822

3.1 Preface

It is crucial for researchers to optimize RNA-seq experimental designs for differential expression detection. Currently, the field lacks general methods to estimate power and sample size for RNA-Seq in complex experimental designs, under the assumption of the negative binomial distribution. We simulate RNA-Seq count data based on parameters estimated from six widely different public datasets and calculate the statistical power in paired and unpaired sample experiments. We comprehensively compare five differential expression analysis packages: DESeq, edgeR, DESeq2, sSeq and EBSeq, and evaluate their performance by power, receiver operator characteristic (ROC) curves and other metrics including Areas Under the Curve (AUC),

Matthews Correlation Coefficients (MCC) and F-measures. DESeq2 and edgeR tend to give the best performance in general. In terms of increasing statistical power, however increasing sample size is more potent than sequencing depth, especially when the sequencing depth reaches 20 million reads. Long intergenic non-coding RNAs (lincRNA) yields lower power relative to the protein coding mRNAs, given their lower expression level in the same RNA-Seq experiment. On the other hand, paired-sample RNA-Seq significantly enhances the statistical power, confirming the importance of considering the multi-factor experimental design. Finally, a local optimal power is achievable for a given budget constraint, and the dominant contributing factor is sample size rather than the sequencing depth. In conclusion, we provide a power analysis tool that captures the dispersion in the data and can serve as a practical reference under the budget constraint of RNA-Seq experiments.

3.2 Introduction

RNA-Seq is a new approach to transcriptome analysis based on next generation sequencing (NGS) technology. It is quickly replacing microarrays as the platform for gene expression profiling, owing to the advantages of high repeatability but low noise level. Beyond revealing gene expression patterns, the information gained from RNA-Seq has already greatly enhanced our understanding in many other areas, such as mechanisms of alternative splicing and the discovery of many novel isoforms of mRNA transcripts [1, 2]. Furthermore, it has led to the discovery of many novel RNA transcripts, as well as the massive amount of newly discovered long intergenic non-coding RNAs (lincRNAs) relative to the small number of lincRNAs identified before RNA-Seq became popula [1–4].

RNA-Seq data are discrete counts, and the Poisson distribution had previously been used to analyze RNA-Seq data [5–10]. Several earlier RNA-Seq studies have attempted to use the Poisson distribution to perform power analysis and sample size estimation using algebraic manipulation of Wald statistics and likelihood ratio methods [11, 12]. Chen et al. studied several test statistics (Wald test, likelihood ratio test, Fishera's exact test, variance stabilized test and conditional binomial test) on Poisson distribution simulations and compared their performances in terms of statistical power [12]. They justified the use of the Poisson distribution in the simulation data by arguing that the Poisson distribution can be used when there are only technical replicates. However, the much larger variation from biologic replicates [13] was not addressed in the paper. Moreover, it was found that the Poisson distribution does not fit the empirical data due to the over-dispersion mainly caused by natural biological variation [8, 14]. As a result, the

negative binomial distribution has become widely used to analyze RNA-Seq data which allows more flexibility in assigning between-sample variation.

It is very challenging to estimate power and satisfactory sample size for the RNA Seq differential expression (DE) tests. One issue is that analytical solutions may not always exist for RNA-Seq sample size and power calculations [7, 15, 16], due to the complexity of the negative binomial model. Instead, numerical methods such as Monte Carlo simulations have been employed to analyze the properties of negative binomial models [9, 15–19]. Other issues involved in power estimation include the combination of multiple hypotheses testing (MHT), p-value calculation, and various ways to estimate dispersion and normalization factors for library sizes. In RNA-Seq analysis, tens of thousands of genes are analyzed for statistical significance simultaneously. A naive approach would analyze each individual gene independently without consideration of the entire dataset. However, since correlations exist among different genes within the same sample as well as the same genes among related samples, more accurate results can be obtained by making sensible assumptions regarding such information. This strategy has been implemented in recent RNA-Seq DE packages such as DESeq, DESeq2, edgeR, EBSeq and sSeq [20–24].

Several studies examined differences between statistical packages of RNA-Seq DE analysis [18, 23, 25, 26]. Nookaew et al. evaluated the differences in DE using the experimental data of yeasts in different growth conditions. Conversely others [18, 23, 25, 27] calculated true positive rate (TPR) and false positive rate (FPR) using simulated datasets under varying parameters. In this report, we took a unique combination of simulated and experimental data approaches, where the parameters in the simulations were based on six different experimental data sets that span a wide range of conditions and samples. This approach is solidly grounded upon realistic RNA-Seq data, yet it is very flexible and can realistically reveal the relationships among parameters relevant to the power analysis. We analyzed the entire simulated datasets as well as sub datasets that are stratified by log2 fold changes (LFC) or expression levels, so that we could detect DE limits given varying parameters in the model. To follow the most recent progress in the RNA-Seq DE area as well as to present results with minimal bias, we selected two widely used methods (DESeq and edgeR) and three recent DE analysis packages released within the past year (DESeq2, EBSeq and sSeq). DESeq2, the most recent derivative of DESeq, was reported to have better power compared to the DESeq package [28]. EBSeq displayed robustness and better performance in analyzing isoform-level expression, yet comparable to other methods in analyzing gene-level expression [21]. Additionally, sSeq package was chosen as it achieved better sensitivity for experiments with small sample sizes [29]. Through comprehensive comparison among all these methods, we aimed to reveal the true relationships between statistical power and its contributing factors.

3.3 Methods

In this study, we evaluated two different types of experimental designs: paired (two-factor) and unpaired (single-factor) designs. In unpaired experimental designs, samples or individuals in one condition are compared to independent samples in another condition. Paired design is a special case of multi-factor (eg. two-factor, three-factor etc.) designs which consider factors that affect the expression level of each sample. Specifically, in paired experiments each sample has two conditions (such as cancer tissue and cancer adjacent normal tissue) that both yield RNA-Seq data. In this study, we used the paired experimental design as a demonstration of the multi-factor design, where the pairing information was treated as the second factor that affects the expression level of each gene.

In our simulated data, we used a general linear model (GLM) with negative binomial distribution. We estimated their parameters from public datasets employed in this study. For the unpaired datasets of two groups, the counts for a particular gene in a sample i were modeled by the formula:

$$log\mu_i = x_i^T + logN_i \tag{1}$$

Here μ_i is the counts for sample i, N_i is the normalized library size for sample i, I is the vector of coefficients for the two different experimental conditions, and x_i is a vector of length two indicating whether sample i belongs to condition one or condition two in the experiment. The LFC was then determined by the difference of the two elements of I. For paired-sample designs, the counts for a gene were modified from (1) with the following formula:

$$log\mu_i = x_{1i}^T\beta + x_{2i}^T + logN_i \tag{2}$$

Here a new vector of coefficients P of length n/2 is introduced to represent the relative expression level for each pair of samples. The other new vector x2i denotes which pair a particular sample belongs to.

The GLM parameters for each gene in each real dataset were estimated by the glm function in R, using a log link function for the count data. The family of negative binomial distributions was calculated by the negative.binomial function in the MASS package. The amount of dispersion per gene was estimated using the Cox-Reid approximate conditional maximum likelihood (CR-APL) method [30] . This method modifies the maximum likelihood estimate of

dispersion by accounting for the experimental design through the Fisher's Information Matrix in the log-likelihood function [30]. CR-APL is implemented as the dispCoxReidInterpolateTagwise function in the edgeR package, and it is also used in DESeq to estimate the dispersion in multi-factor experimental designs.

3.3.1 Generation of simulated count data

The count data were generated from the negative binomial distribution. For each gene, the count Yi was given by:

$$Y_i \sim NB(mean = \mu_i, var = \mu_i(1 + \mu_i\phi_i)) \tag{3}$$

Here, ϕ_i is the per-gene dispersion calculated by the CR-APL method, and the expected value μ_i is a function of the library size. The library size of each simulated sample was generated from a uniform distribution whose parameters were estimated from the maximum and minimum of the real dataset.

We used five statistical packages for DE testing: DESeq (version 1.14.0) and edgeR (version 3.4.2) methods, as well as three newer packages released within the past year: DESeq2 (version 1.2.9), EBSeq (version 1.3.1) and sSeq (version 1.0.0). All packages are implemented in the Bioconductor/R platform. We determined the truth data for DE in the simulation as the overlapping DE genes detected from all five statistical packages used in the study, using the original real datasets. This approach is similar to other studies [23, 27]. In the simulation, the LFC of DE genes was determined by the equation 1 and 2. We set the LFC of genes that are not differentially expressed to zero in the generation of the simulated count data, as done by others [27].

3.3.2 Description of public datasets used in the study

The six public datasets are listed in Table 1 (see Results). We enumerate the parameters of each dataset in the following:

Bottomly – We used this published data set to compare gene expression between C57BL/6J and DBA/2J mouse strains [31]. An average of 22 million reads was generated for 21 mice (10 C57BL/6J and 11 DBA/2J). Count data were downloaded from the ReCount project [32].

Bullard – Ambion's human brain reference RNA (brain) and Stratagene's Universal Human Reference(UHR) RNA were compared [33]. An average of 12.5 million reads was generated from 7 brain and 7 UHR technical replicates. Count data were also downloaded from ReCount project [32].

Huang a- Differentiated embryonic stem cells were compared with fetal head tissues of 14.5 days post coitum. Four biological samples were compared using various rRNA removal methods, in order to analyze coding and non-coding RNAs [34]. Twenty-two technical replicates were used with an average of 17.7 million reads per sample. Short Read Archive (SRA) reads were downloaded from GEO (GSE22959) and aligned with tophat to mm10 reference genome. Count data were generated using HTSeq [35].

Montgomery-Pickrell (M-P) – RNA-seq data from 60 individuals of European descent [36] and 69 individuals of Nigerian descent [37] were sequenced with an average sequencing depth of 17 million reads per sample. The datasets were used to analyze DE between the two populations. Count data were downloaded from the ReCount project [32].

Tuch – Three paired tumor and non-tumor tissues from oral squamous cell carcinoma patients were sequenced for an average of 205 million reads per sample [38]. Count data were downloaded from Table S1 of the original publication.

Qian – West Nile Virus (WNV) transfection of macrophage cells from 10 healthy donors were compared to mock transfection of the same cell culture with a total of 28 million reads per sample [39]. Raw SRA read data were downloaded from GEO (GSE40718) and aligned with tophat to Hg19 Refseq genes downloaded from UCSC Genome Browser. Count data were also generated using HTSeq.

3.3.3 Detection of DE in unpaired (single-factor) experimental designs

We aimed to calculate p-values, sensitivity (power) and specificity over the range of parameters. Toward these aims, we performed standard analyses with functions implemented in the five RNA-Seq analysis packages DESeq2, DESeq, edgeR, sSeq and EBSeq. Specifically, in DESeq the count data were analyzed using *newCountDataSet*, followed by *estimateSizeFactors*, *estimateDispersions* and nbinomTest functions. For *DESeq2*, *DESeqDataSetFromMatrix* was used, followed by *estimateSizeFactors*, *estimateDispersions* and *nbinomWaldTest* functions. For edgeR, count data were analyzed using *DGEList* followed by *calcNormFactors*, *estimateCommonDisp*, *estimateTagwiseDisp* and *exactTest* functions. For EBSeq, the libraries were first normalized using *MedianNorm* and then DE genes were detected using the *EBTest* function. For sSeq,

DE genes were detected using *nbTestSH* function. In packages where p-value adjustment was needed, the p.adjust function in R with method="BH" (Benjamini and Hochberg FDR option) was employed.

3.3.4 Detection of DE in paired-sample (two-factor) experimental designs

Similar to the unpaired or single-factor designs, we performed standard analyses for the paired-sample experimental designs. We calculated p-values, sensitivity (power) and specificity over the range of parameters, using four statistical packages: DESeq2, DESeq, edgeR and sSeq. We did not conduct DE gene detection using EBSeq, as it is not adapted to analyzing paired data currently (communication with the authors). In DESeq, data were analyzed similar to above, using the two-factor design matrix and *method="pooled-CR"* for the dispersion estimation, followed by *fitNbinomGLMs* function for both the null and alternative hypotheses, and then by *nbinomGLMTest* function to calculate p-values per gene. DESeq2 was used similarly to the single-factor analysis above, using the two-factor design matrix (condition + pairing information). For edgeR, count data were analyzed using *DGEList* followed by *calcNormFactors*, *estimateCommonDisp*, *estimateGLMTrendedDisp*, *estimateTagwiseDisp*, *glmFit* and *glmLRT* functions. For sSeq, function *nbTestSH* was used with *pairedDesign=TRUE*. P-value adjustment was done the same way as in single-factor design, when needed.

3.3.5 Calculation of true positive rates (power) and false positive rates

The sample sizes in the simulated data sets varied from n=2 to n=25 and the average library sizes varied from 1 million up to 50 million reads. Each condition was simulated 100 times using random seeds uniformly distributed from 1 to 100 as the inputs. Given a significance threshold of 0.05, the TPR was calculated by:

$$TPR\ (power) = \frac{TP}{TP + FN} \tag{4}$$

And the FPR was calculated by:

$$FPR = 1 - Specificity = \frac{FP}{FP + TN} \tag{5}$$

Two standard performance measures, Matthews correlation coefficient (MCC, also known as the phi statistic) and F-measure are calculated by:

$$MCC = \frac{TP * TN - FP * FN}{\sqrt{(TP + FP)(TP + FP)(TN + FP)(TN + FN)}} \tag{6}$$

and:

$$F\_measure = \frac{2 * TP/(TP + FP) * TP/(TP + FN)}{\frac{TP}{TP+FP} + \frac{TP}{TP+FN}} \tag{7}$$

Where TP = true positives, TN = true negatives, FP = false positives, and FN = false negatives.

3.3.6 Planning RNA-Seq under the budget constraint

For RNA-Seq cost calculation, we referred to Illumina Hi-Seq single-end RNA-Seq prices listed by the Yale Center for Genome Analysis (`http://ycga.yale.edu/services/illuminaprices.aspx`). The total overhead cost of each sample was estimated as $241, which includes sample quality check and mRNA library construction. The remaining sequencing cost per lane was $1331 based on HiSeq 2000 single-end sequencing. Simulated count data were generated as before, by modelling gene counts through equation (1) or (2) and the negative binomial distribution. The formula to calculate the budget is as follows:

$$\begin{aligned} Budget = {} & overhead\ cost\ per\ sample * number\ of\ samples \\ & + sequencing\ depth\ per\ sample * number\ of\ samples \\ & / sequencing\ depth\ per\ lane * cost\ per\ lane \end{aligned} \tag{8}$$

All R code is available for downloading from our website: `http://www2.hawaii.edu/~lgarmire/RNASeqPowerCalculator.htm`

3.4 Results

3.4.1 Estimation of parameters in the datasets

We based our simulation results on six representative RNA-Seq datasets. The description of parameters for these datasets is summarized in Table 1. Among them, four datasets used polyA enriched method while the other two used Ribosome depletion method. Four datasets had unpaired experimental designs and two had paired-sample designs. The datasets have a wide

variety of sample sizes ranging from 6 samples (Tuch) to 129 samples (M-P), as well as a variety of experimental conditions spanning from cell-line, tissue, viral infection, cancer to population comparisons. We chose this variety to capture the wide range of different parameters from various types of experiments.

We estimated the parameters from each of the six datasets and fit them by GLMs with negative binomial distribution (Table 1). For unpaired data sets, we used the five RNA-Seq analysis packages (DESeq, DESeq2, edgeR, EBSeq and sSeq) to detect DE genes, whereas for paired data sets, EBSeq was not used as it is not adapted to the paired-design (communications with authors). We took a conservative approach to call DE genes by taking intersected DE genes from all four or five RNA-Seq analysis packages (Table 2).

In summary, the library sizes (reads mapped to the transcriptome) of the six datasets range from a log10 mean of 6.15 (Bullard) to 7.43 (Qian), the normalized median gene expressions log2 counts per million (CPM) ranged from 3.18 (Huang) to 4.61 (Bullard), and the median LFCs of DE genes range from 3.33 (Huang) to 0.751 (M-P). Among them, the Bullard dataset which compared between brain tissue and the UHR RNA library had the highest percent DE (59.3%) and a median LFC (2.13). The samples for this dataset were technical replicates and thus the median dispersion is extremely low (0.000391). In contrast, the M-P data that compared Caucasian to African populations have a much lower percent DE (21.5%) and the highest median dispersion (0.231). These results indicate that comparisons at tissue levels (eg. Huang and Bullard) have more significant differences between conditions, whereas comparisons at the population level (eg. M-P) have a very small significant change due to the large heterogeneity among populations.

3.4.2 Effects of experimental parameters on power of RNA-Seq analysis

Due to the cost of RNA-Seq experiments, it is imperative to know prior to an experiment the number of biological replicates required to achieve the desirable power among genes of interest (e.g., specific expression levels and/or fold change range). We used the negative binomial distribution and approximate parameters from six RNA-Seq public datasets to create simulated data for unpaired and paired experiments. We performed 100 simulations per condition to calculate the statistical power for five categories of DE genes: all DE genes, DE genes with low expression, high expression, low fold change (FC) and high FC that are separated by quartiles.

Fig. 1 and S1 show the comparisons among the six data sets, five DE categories and five DE detection methods. We observed the following patterns: (1) In general, higher power is achieved

as the number of replicates increases. However, beyond a certain replicate number, the gain in power gain is negligible. This saturation replicate number is dependent on dispersion and median LFC of the data: the smaller the dispersion or the bigger median LFC, the smaller the number of replicates is to reach saturation. However, EBSeq showed lower power at higher replicates in the subset of genes with high expression for the Huang dataset, potentially due to a problem to handle large counts in the simulation (communications with authors); (2) higher power is achieved as the sequencing depth increases, however beyond 5-20 million reads, depending on the data set, the gain in power gain is minimal (Fig. S1). Similar to sample size, the smaller the dispersion or the bigger median LFC (Bullard and Tuch data), the smaller the sequencing depth is to reach saturation; (3) High FC and high gene expression quartiles generally show increased power over low LC and low gene expression quartiles, before the saturation point of replicates; (4) Power is highly affected by the experimental conditions, and (5) No single DE program shows consistently the highest power across all datasets. The relationships among power, sample size and datasets are complicated. However, some general trends emerge: when the dispersions are small (Bottomly, Bullard, M-P, Qian and Tuch data), edgeR and DESeq2 generally give higher power estimations, especially when the replicates i $\leq$ 5. However when the dispersion is large (M-P data), sSeq yields the highest power. Generally, DESeq estimates power more conservatively, confirming the results of Robles et al. [18].

Given that power is highly dependent on the dataset, we examined the relationships among power, dispersion, and sample size further (Fig. 1 and Table 1). Simulations based on the Bullard and Tuch data show that all programs achieve very high power close to 1 (e.g., 100% detection of DE genes). In the Bullard dataset, the estimated median dispersion parameter is extremely low (0.000391). This is likely due to that fact that the samples in this dataset are technical replicates rather than biological replicates. Thus all DE analysis packages used here could easily detect differences between the two groups. In the Tuch data, the high power was achieved largely due to the high median FC of DE genes (2.13, Table 1) and pair-designed samples. On the opposite side, the M-P data consist of transcriptomes from 129 individuals. The M-P dataset had the highest median dispersion of 0.231 and the lowest median FC=0.751 (Table 1). Only DESeq2, edgeR and sSeq were able to achieve a power of 0.8 or greater at a sample size of 25 replicates per condition (Fig. 1).

3.4.3 Performance analysis of other metrics

In addition to statistical power (sensitivity), specificity (complement of FPR) is also an important factor to assess the performance of each DE program. To evaluate them together, we

generated Receiver Operator Characteristic (ROC) curves based on the results of the simulated data with 4 replicates per condition (Fig. 2A). The most optimal ROC curve jointly displays high levels of TPR and FPR. DESeq2 and edgeR had similar and the best ROC curves for all datasets. DESeq performed similarly to DESeq2 and edgeR, except for the M-P data. However, EBSeq and sSeq generally did not perform as well as the others. EBSeq sometimes yields a large increase in FPR with little corresponding increase in TPR, suggesting its limitation to control the type I error. We also evaluated the different programs with other performance metrics: Area Under the Curve (AUC) of ROC curves, Matthews Correlation Coefficient (MCC) which takes into account all true and false positives and negatives, and F-measure which is the weighted average of the precision and sensitivity (Fig. 2B). Although we see that no single package consistently performs the worst or the best in all data sets, we did observe similar results as in the ROC curves: DESeq2 and edgeR generally have similar and the best AUC, MCC and F-measure, except for the M-P data in which sSeq has the best MCC and F-measure (Fig. 2B, Fig. S2 and S3; Table S1).

3.4.4 Improved statistical power by the paired-sample design

In the experimental design, multiple conditions or factors can be set up to affect the expression level of each biological sample. For example, in paired-design experiments, each biological sample has two conditions (such as cancer tissue and cancer adjacent normal tissue) to generate RNA-Seq data. In this study, we used the paired-sample design as a demonstration of multi-factor design, and treated the pairing information as the second factor that affects the expression level of each gene. We used a GLM with negative binomial distribution to estimate the effects of the experimental condition and pairing information, based on parameters estimated from the two paired datasets (Qian and Tuch data). Fig. 3 shows the comparisons among the four DE categories in these two datasets, under either single-factor (unpaired) or paired statistical model. It is clear that by considering the pairing information, the statistical power is increased, especially for the Qian data. The Qian dataset has a lower median LFC (0.929) relative to the Tuch dataset (2.13), as well as a lower median dispersion (roughly 40% lower than Tuch). This suggests a big advantage to better differentiate genes by introducing additional pairing restrictions, when the overall LFC among genes is not very large. Similar to Fig. 1 and regardless of single-factor or paired-sample model, we observed that DESeq2 and edgeR give the highest power estimations when the number of replicates is small; however sSeq quickly catches up when the number of replicates increases. Again DESeq gives the most conservative estimation of power among the four DE test methods.

3.4.5 Differences in experimental power based on transcript type

Depending on the subsets of transcripts of interest, there might be differences for achievable power. For example, lincRNA are generally expressed at low or medium levels relative to mRNAs from protein coding genes [40, 41]. Thus the mRNA transcriptome and lincRNA transcriptome may yield different levels of power, even when they are generated from the same RNA-Seq experiments and the same biological samples. To test this, we conducted simulations based on the Huang dataset. This dataset was chosen because it used ribosomal RNA depletion method rather than poly-A selection, so that lincRNA detection was enhanced. To show the internal difference of the two types of RNAs, we divided the dataset by the type of transcripts and summarized the parameters (Table 3). Indeed the most striking difference between the two types of RNAs is the median expression level: the mRNA has a median expression measured in log2 CPM of 4.63, whereas the lincRNA only has a median log2 CPM of 1.25. As expected, the analysis of protein coding genes had higher power compared to the analysis of lincRNA transcripts when the number of replicates i$\leq$3, which is often the limit for many experimental labs (Fig. 4). DESeq is most conservative in power estimation and showed the largest difference in power between the two types of transcripts, especially when the number of replicates is low. At four replicates per condition, lincRNAs had a power of 0.65 compared to protein coding genes power of 0.75. However, when the number of replicates is sufficient, this difference of power becomes minimal.

3.4.6 Optimize sample size and sequencing depth under the budget constraint

In real-world RNA-Seq experimental design, the budget constraints usually exist and can significantly affect the trade-off decision between the sample size and sequencing depth. To demonstrate the practical application of RNA-Seq power analysis, we conducted 100 simulations per condition to approximate the optimal sample size and sequencing depth, exemplified by several different budget constraint scenarios ($3000, $5000, $10,000). The cost of RNA-Seq per sample is dependent on the cost of constructing the RNA-Seq library, as well as the cost of sequencing depth (or library size) per sample under the multiplex arrangement, where multiple samples will be barcoded to share one lane of the flow cell. We used an estimated cost of $241 for library construction and $1331 for single-end sequencing cost per lane. Since that not all reads map to the transcriptome, we used a mapping percentage of 20%. We determined the optimal power, corresponding sample size and sequencing depth based on the parameters estimated from the six datasets (Fig. 5 and S4). As demonstrated by the Bottomly data in Fig. 5, the higher

the budget cap is, the more biological replicates are needed (Fig. 5 A and C) to reach the optimal power (Fig. 5 A and B); however the sequencing depth does not change much relative to biological replicates and stays around 20 Million, estimated from most DE methods (Fig. 5D). The highest power was achieved by sSeq, followed closely by DESeq2 and edgeR (Fig. 5 A and B). However, sSeq also showed larger standard deviations in the estimated power compared to the other programs (Fig. 5A). DESeq, DESeq2 and edgeR tend to give rise to less skewed power curves across number of replicates, relative to EBSeq and sSeq (Fig. 5A). EBSeq tends to yield lower optimal power estimation and skews towards fewer replicates but higher sequencing depths, whereas sSeq favors more replicates and lower sequencing depth (Fig. 5 A and D).

3.5 Discussion

RNA-Seq technology is gradually replacing microarray as the method to detect transcriptome level gene expression, therefore it is a critical time to address the problem of desirable statistical power in the RNA-Seq experimental design. There have been a few papers on power and sample size estimation in RNA-Seq experiments; however, these methods need improvement to capture the dispersion in the data and serve as a practical guideline given budget constraints. Busby et al. (Busby et al. 2013) measured power as the percentage of genes with 2-fold count change (by default) that were correctly detected based on the statistical t-test, without realistically capturing the underlying data structure. Hart et al. performed analysis on 127 RNA-Seq samples in human versus fish 2 organisms [42]. They derived a first order closed form approximation of GLM to compute required sample size and desired power, by taking into account of the variance, expected expression level and fold change. Alternatively, Li et al. proposed an exact test to replace hyper-geometric probabilities with the negative binomial distribution [43]. However, neither of their methods considered these complexities: (1) more complicated multi-factor experimental designs, (2) the various ways to estimate dispersion through different analysis packages (they only used edgeR package), and (3) practical optimization of experimental design given a budget cap. The trade-off between sequencing depth and the number of biological samples was recently studied [44]. The authors discovered that adding biological replicates increases the power to detect DE genes better than the strategy of increasing sequencing depth. However, they did not provide a direct solution for optimization given the fixed budget. Moreover like the others, they did not consider multiple DE analysis packages, multi-factor experimental design, or large scale RNA-Seq experiments such as in the population-based studies.

Compared to these earlier studies, we have made a major leap-forward, rather than incremental progress towards providing first-hand and comprehensive references in consideration of RNA-Seq experimental design. We systematically evaluated five popular or more recent DE packages, and conducted simulations based on 212 RNA-Seq samples from six different datasets that span a wide range of experimental conditions, from cell-line, tissue, viral infection, cancer and population comparisons. We chose the truth data based on more coherent criterion, the intersection of DE genes that are consistent from all different RNA-Seq analysis packages, rather than the more arbitrary LFC threshold like others (Robles et al. 2012; Kvam et al. 2012). Moreover, we provided a reference framework to analyze paired-sample, or more general multi-factor experiments, using the GLM approach. Last but not least, we have provided a tool to enable researchers to determine the sample size that optimizes the power, when the budget is limited.

Our study provides many aspects of practical guidance towards the RNA-Seq experimental design. First, dispersion shows a striking impact on power. In datasets with very low dispersions, such as the Bullard data, a power of 0.8 is easily reached with very low sample size and sequencing depth. On the other hand, in datasets with high dispersion, such as M-P data, a power of 0.8 is hardly achievable except at the highest limits of simulation parameters. Due to the strong effect of dispersion, it is clear that statistical tests based on the Poisson distribution (i.e., assuming dispersion = 0) are not capable of handling situations with significant biological variation. Dispersion is primarily due to biological variation, however it can also be attributed from technical variability such as lane differences and the "shot noise" of the random process [13]. Genes with lower expression have high variance [14], and the subset of DE genes in this group are more likely to have higher fold change [45]. All of these factors lead to the challenge of proper estimation of dispersions in the RNA-Seq experiments.

Different RNA-Seq DE testing packages estimate dispersion differently, making the systematical comparisons of these packages worthwhile. We compared the power and other metrics, such as AUC of the ROC curve, MCC and F-measures in five popular or most recent packages. For most datasets, DESeq2 and edgeR give the highest estimate of power, closely followed by DESeq (except the M-P data). DESeq (by default) estimates dispersion by pooling all samples together, fitting them to a parametric distribution and conservatively taking the maximum. This conservative approach may explain why DESeq gives a relatively lower power, as also noted by others [18]. DESeq2 is the new update to DESeq, and it uses shrinkage estimation for dispersion: the first round of dispersion-mean relationship is obtained by maximum likelihood estimates (MLE), and this fit is then used as a prior to estimate the maximum a posteriori estimate for dispersion in the second round. EdgeR estimates dispersion differently, it moderates the dispersion per gene towards a common value across all genes, or towards a local estimate with

genes of similar expression. For paired-sample designs, the DESeq package recommends using the Cox-Reid approximate conditional maximum likelihood (CR-APL) method [20]. DESeq2 likewise uses the CR-APL method to derive dispersion per gene, and then shrinks the dispersion towards a parametric fit assuming a prior distribution of log dispersion [46]. edgeR also uses CR-APL and then shrinks the dispersion estimate using empirical Bayes [24]. On the other hand, EBseq estimates dispersion by the method of moments, and then uses Bayes posterior probabilities as the measure of statistical significance. While EBSeq generally does not perform as well as other packages, it could outperform others on analyzing isoform level expression [21], rather than gene level expression which is the focus of this report. sSeq estimates dispersion by pooling all the samples together using the method of moments, and then shrinking the per-gene estimates through minimizing the mean-square error [23]. Although the authors of sSeq stated that sSeq compared favorably to other popular packages in low sample sizes regarding sensitivities and specificities, using an external gold standard [29], we found that it did not yield the highest powers in the Bottomly and Tuch datasets when the replicates are i$\leq$5. This indicates that the performance of sSeq is affected by the data sets or the choice of truth measure.

Two other important factors that influence power are the number of replicates and sequencing depth. In general, more biological replicates and greater sequencing depth help to achieve greater statistical power to a certain extent. Sequencing depth is closely related to the expected counts of genes. As sequencing depth increases within the range of 5-20 million reads, genes with lower expression levels, lower fold change and higher dispersions become detectable [47]. However, above 20 million reads, the contribution of sequencing depth to power gain becomes minimal. Combined with preliminary data, sequencing depth can be used for investigating genes of certain expression strengths. For example, if one were interested in estimating the statistical power for lincRNAs, which are on average transcribe 10-fold lower than mRNA transcripts [40], one would not be as concerned about the FDR adjustment for the entire dataset. It is therefore possible to enumerate the power and sample size for transcripts of a specific type (e.g., genes with low versus high expression) or over a certain range of parameters (e.g., low LFC versus high LFC). Based on our results, we would recommend a minimum of 5 replicates in order to diminish the power difference between protein coding mRNA and lincRNAs for the sequencing depth of around 20 million reads.

We also aimed to generalize the potential uses of two-factor analysis by estimating parameters from two paired-sample datasets: The Tuch dataset is a paired cancer and normal tissue experiment, and the Qian dataset is a paired West Nile Virus and mock transfection of cell cultures. We compared the power to detect DE genes in these two sets using paired analysis versus one-factor analysis, and showed that two-factor models can substantially increase detection limit

and hence power in RNA-seq analysis. Furthermore, DESeq, DESeq2 and edgeR are capable of arbitrary design matrices, including scenarios such as time series and blocking design that reduces known variability in confounding factors.

We demonstrated the optimization of RNA-Seq experiments under the budget constraint, a real-world problem for investigators. We showed that a local optimum of power is achievable for a particular samples size. More importantly, we found that the dominant contributing factor to reach optimal power at specific a budget constraint is sample size, rather than sequencing depth which is around the 20M reads range for most DE detection packages. This conclusion is consistent with Liu et al [44], in that biological replicates are more important than read depth for DE detection, although we investigated differently from the power perspective with budget constraints. DESeq, DESeq2 and edgeR presented more symmetrical curves of sample size versus power, whereas EBSeq and sSeq seemed to be more skewed. Correlating to the ROC curves and earlier power estimation without budget constraints, DESeq2 and edgeR appear to be the better choices of software for their overall performances.

As RNA-Seq technology matures and sequencing becomes cheaper, complex experiments with more replicates and greater sequencing depth will become more prevalent and there will be an increasing need to design RNA-Seq experiments more thoughtfully. Our approach reported here can be applied more generally to complex multi-factor designs that can be modelled through the GLM framework, such as time series, multi-level designs and blocking designs. We have also demonstrated how optimal sample size and power can be calculated, given a budget constraint. It is our expectation that researchers will find our methods useful and valuable in designing RNA-Seq differential expression experiments.

3.6 Acknowledgements

The authors thank Dr. Lynne Wilkens for reviewing the manuscript, and Dr. Christina Kendziorski and her students for the communications in using EBSeq. The authors are also grateful to Dr. Gordon Okimoto and Mr. Mike Loomis for providing access to the server clusters of Biostatistics and Bioinformatics Shared Resources at University of Hawaii Cancer Center.

References

1. Kim, D. & Salzberg, S. L. TopHat-Fusion: an algorithm for discovery of novel fusion transcripts. *Genome Biology* **12.** ISSN: 1465-6906. doi:10.1186/gb-2011-12-8-r72 (2011).
2. Morin, R. D., Bainbridge, M., Fejes, A., Hirst, M., Krzywinski, M., Pugh, T. J., McDonald, H., Varhol, R., Jones, S. J. M. & Marra, M. A. Profiling the HeLa S3 transcriptome using randomly primed cDNA and massively parallel short-read sequencing. *Biotechniques* **45** (2008).
3. Trapnell, C., Pachter, L. & Salzberg, S. L. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* **25,** 1105–1111. ISSN: 1367-4803, 1460-2059 (2009).
4. Wang, Z., Gerstein, M. & Snyder, M. RNA-Seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics* **10,** 57–63. ISSN: 1471-0056 (2009).
5. Jiang, H. & Wong, W. H. Statistical inferences for isoform expression in RNA-Seq. *Bioinformatics* **25,** 1026–1032 (2009).
6. Marioni, J. C., Mason, C. E., Mane, S. M., Stephens, M. & Gilad, Y. RNA-seq: An assessment of technical reproducibility and comparison with gene expression arrays. *Genome Research* **18,** 1509–1517. ISSN: 1088-9051, 1549-5469 (2008).
7. Pham, T. V. & Jimenez, C. R. An accurate paired sample test for count data. *Bioinformatics* **28,** i596–i602 (2012).
8. Robinson, M. D. & Oshlack, A. A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol* **11** (2010).
9. Srivastava, S. & Chen, L. A two-parameter generalized Poisson model to improve the analysis of RNA-seq data. *Nucleic Acids Research* **38,** e170–e170 (2010).
10. Wang, L., Feng, Z., Wang, X., Wang, X. & Zhang, X. DEGseq: an R package for identifying differentially expressed genes from RNA-seq data. *Bioinformatics* **26,** 136–138 (2010).
11. Busby, M. A., Stewart, C., Miller, C. A., Grzeda, K. R. & Marth, G. T. Scotty: a web tool for designing RNA-Seq experiments to measure differential gene expression. *Bioinformatics* **29,** 656–657 (2013).
12. Chen, Z., Liu, J., Ng, H. K. T., Nadarajah, S., Kaufman, H. L., Yang, J. Y. & Deng, Y. Statistical methods on detecting differentially expressed genes for RNA-seq data. *BMC Systems Biology* **5** (2011).
13. McIntyre, L. M., Lopiano, K. K., Morse, A. M., Amin, V., Oberg, A. L., Young, L. J. & Nuzhdin, S. V. RNA-seq: technical variability and sampling. *BMC genomics* **12** (2011).

14. Anders, S. & Huber, W. Differential expression analysis for sequence count data. *Genome Biol* **11** (2010).

15. Aban, I. B., Cutter, G. R. & Mavinga, N. Inferences and Power Analysis Concerning Two Negative Binomial Distributions with An Application to MRI Lesion Counts Data. *Computational statistics & data analysis* **53,** 820–833. ISSN: 0167-9473 (2008).

16. McCulloch, C. E. Maximum Likelihood Algorithms for Generalized Linear Mixed Models. *Journal of the American Statistical Association* **92.** ISSN: 01621459. doi:10.2307/2291460 (1997).

17. Robinson, M. D. & Smyth, G. K. Moderated statistical tests for assessing differences in tag abundance. *Bioinformatics* **23,** 2881–2887. ISSN: 1367-4803, 1460-2059 (2007).

18. Robles, J. A., Qureshi, S. E., Stephen, S. J., Wilson, S. R., Burden, C. J. & Taylor, J. M. Efficient experimental design and analysis strategies for the detection of differential expression using RNA-Sequencing. *BMC genomics* **13** (2012).

19. Vijay, N., Poelstra, J. W., Kunstner, A. & Wolf, J. B. W. Challenges and strategies in transcriptome assembly and differential gene expression quantification. A comprehensive in silico assessment of RNA-seq experiments. *Molecular Ecology* **22,** 620–634. ISSN: 1365-294X (2013).

20. Anders, S. Analysing RNA-Seq data with the DESeq package. *Mol Biol,* 1–17 (2010).

21. Leng, N., Dawson, J. A., Thomson, J. A., Ruotti, V., Rissman, A. I., Smits, B. M., Haag, J. D., Gould, M. N., Stewart, R. M. & Kendziorski, C. EBSeq: an empirical bayes hierarchical model for inference in RNA-seq experiments. *Bioinformatics* **29,** 1035–1043. ISSN: 1367-4803 (2013).

22. Love, M., Anders, S. & Huber, W. Differential analysis of RNA-Seq data at the gene level using the DESeq2 package. *Bioconductor* (2013).

23. Rapaport, F., Khanin, R., Liang, Y., Pirun, M., Krek, A., Zumbo, P., Mason, C. E., Socci, N. D. & Betel, D. Comprehensive evaluation of differential gene expression analysis methods for RNA-seq data. *Genome Biology* **14,** R95. ISSN: 1465-6906 (2013).

24. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26,** 139–140 (2010).

25. Kvam, V. M., Liu, P. & Si, Y. A comparison of statistical methods for detecting differentially expressed genes from RNA-seq data. *American journal of botany* **99,** 248–256 (2012).

26. Nookaew, I., Papini, M., Pornputtapong, N., Scalcinati, G., Fagerberg, L., UhlA(C)n, M. & Nielsen, J. A comprehensive comparison of RNA-Seq-based transcriptome analysis from reads to differential gene expression and cross-comparison with microarrays: a case study in Saccharomyces cerevisiae. *Nucleic Acids Research* **40,** 10084–10097 (2012).

27. Soneson, C. & Delorenzi, M. A comparison of methods for differential expression analysis of RNA-seq data. *BMC bioinformatics* **14,** 91. ISSN: 1471-2105 (2013).

28. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-Seq data with DESeq2. *bioRxiv* (2014).

29. Yu, D., Huber, W. & Vitek, O. Shrinkage estimation of dispersion in Negative Binomial models for RNA-seq experiments with small sample size. *Bioinformatics* **29,** 1275–82. ISSN: 1367-4803 (2013).

30. McCarthy, D. J., Chen, Y. & Smyth, G. K. Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. *Nucleic Acids Research* **40,** 4288–4297. ISSN: 0305-1048, 1362-4962 (2012).

31. Bottomly, D., Walter, N. A., Hunter, J. E., Darakjian, P., Kawane, S., Buck, K. J., Searles, R. P., Mooney, M., McWeeney, S. K. & Hitzemann, R. Evaluating gene expression in C57BL/6J and DBA/2J mouse striatum using RNA-Seq and microarrays. *PLoS One* **6,** e17820. ISSN: 1932-6203 (2011).

32. Frazee, A., Langmead, B. & Leek, J. ReCount: A multi-experiment resource of analysis-ready RNA-seq gene count datasets. *BMC bioinformatics* **12,** 449. ISSN: 1471-2105 (2011).

33. Bullard, J. H., Purdom, E., Hansen, K. D. & Dudoit, S. Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. *BMC bioinformatics* **11.** ISSN: 1471-2105. doi:10.1186/1471-2105-11-94 (2010).

34. Huang, R., Jaritz, M., Guenzl, P., Vlatkovic, I., Sommer, A., Tamir, I. M., Marks, H., Klampfl, T., Kralovics, R. & Stunnenberg, H. G. An RNA-Seq strategy to detect the complete coding and non-coding transcriptome including full-length imprinted macro ncRNAs. *PLoS One* **6,** e27288. ISSN: 1932-6203 (2011).

35. Anders, S. HTSeq: Analysing high-throughput sequencing data with Python. *Bioinformatics* (2010).

36. Montgomery, S. B., Sammeth, M., Gutierrez-Arcelus, M., Lach, R. P., Ingle, C., Nisbett, J., Guigo, R. & Dermitzakis, E. T. Transcriptome genetics using second generation sequencing in a Caucasian population. *Nature* **464,** 773–7. ISSN: 0028-0836 (2010).

37. Pickrell, J. K., Marioni, J. C., Pai, A. A., Degner, J. F., Engelhardt, B. E., Nkadori, E., Veyrieras, J.-B., Stephens, M., Gilad, Y. & Pritchard, J. K. Understanding mechanisms underlying human gene expression variation with RNA sequencing. *Nature* **464,** 768–772 (2010).

38. Tuch, B. B., Laborde, R. R., Xu, X., Gu, J., Chung, C. B., Monighetti, C. K., Stanley, S. J., Olsen, K. D., Kasperbauer, J. L. & Moore, E. J. Tumor transcriptome sequencing reveals allelic expression imbalances associated with copy number alterations. *PLoS One* **5** (2010).

39. Qian, F., Chung, L., Zheng, W., Bruno, V., Alexander, R. P., Wang, Z., Wang, X., Kurscheid, S., Zhao, H. & Fikrig, E. Identification of genes critical for resistance to infection by West Nile virus using RNA-Seq analysis. *Viruses* **5,** 1664 (2013).

40. Cabili, M. N., Trapnell, C., Goff, L., Koziol, M., Tazon-Vega, B., Regev, A. & Rinn, J. L. Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. *Genes & Development* **25,** 1915–1927. ISSN: 0890-9369, 1549-5477 (2011).

41. Garmire, L. X., Garmire, D. G., Huang, W., Yao, J., Glass, C. K. & Subramaniam, S. A global clustering algorithm to identify long intergenic non-coding RNA-with applications in mouse macrophages. *PLoS One* **6,** e24051. ISSN: 1932-6203 (2011).

42. Hart, S. N., Therneau, T. M., Zhang, Y., Poland, G. A. & Kocher, J.-P. Calculating Sample Size Estimates for RNA Sequencing Data. *Journal of Computational Biology* **20,** 970–978. ISSN: 1066-5277 (2013).

43. Li, C. I., Su, P. F. & Shyr, Y. Sample size calculation based on exact test for assessing differential expression analysis in RNA-seq data. *BMC bioinformatics* **14,** 357. ISSN: 1471-2105 (Electronic) 1471-2105 (Linking) (2013).

44. Liu, Y., Zhou, J. & White, K. P. RNA-seq differential expression studies: more sequence or more replication? *Bioinformatics* **30,** 301–4. ISSN: 1367-4811 (Electronic) 1367-4803 (Linking) (2014).

45. Mutch, D. M., Berger, A., Mansourian, R., Rytz, A. & Roberts, M.-A. The limit fold change model: a practical approach for selecting differentially expressed genes from microarray data. *BMC bioinformatics* **3,** 17. ISSN: 1471-2105 (2002).

46. Wu, H., Wang, C. & Wu, Z. A new shrinkage estimator for dispersion improves differential expression detection in RNA-seq data. *Biostatistics* **14,** 232–243. ISSN: 1465-4644 (2013).

47. Tarazona, S., Garcia-Alcalde, F., Dopazo, J.-n., Ferrer, A. & Conesa, A. Differential expression in RNA-seq: a matter of depth. *Genome Research* **21,** 2213–2223 (2011).

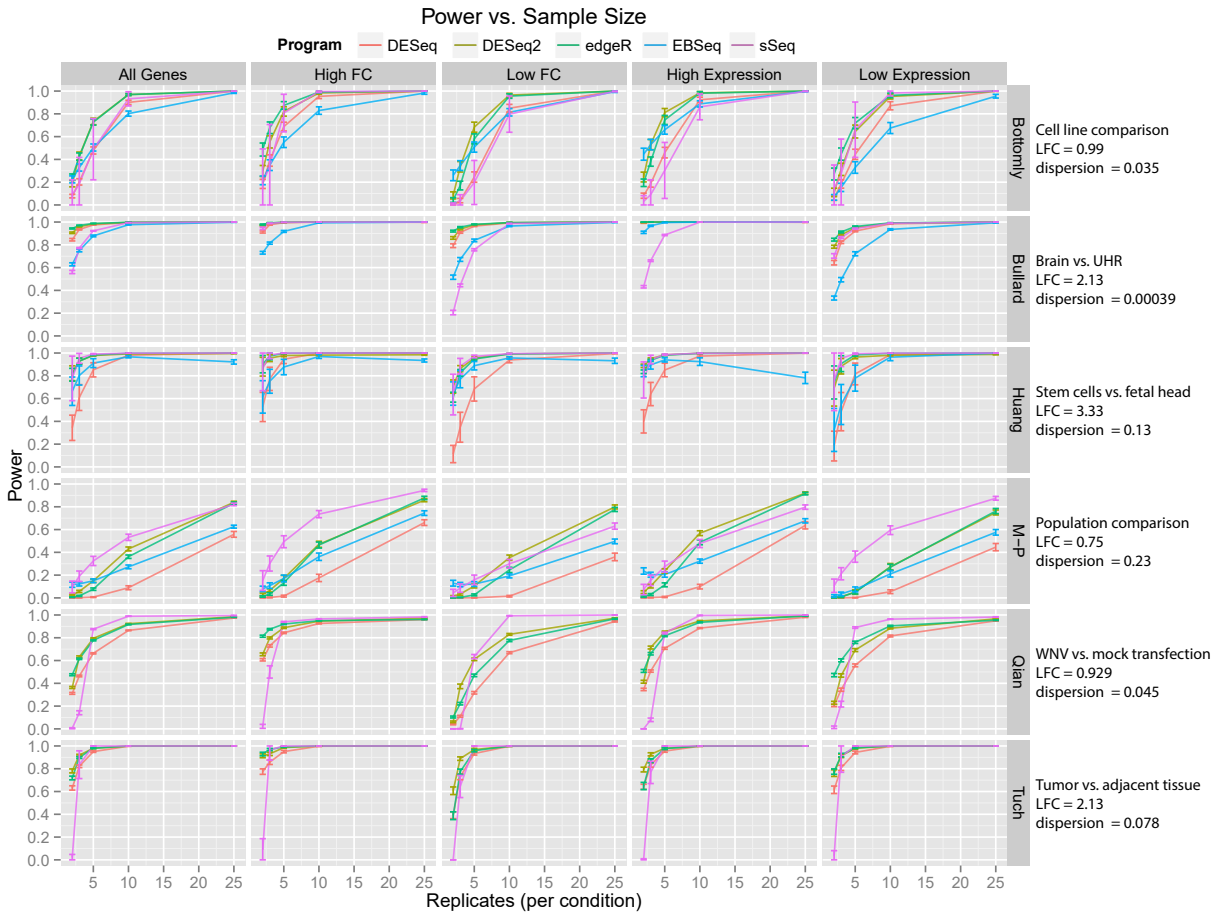


Figure 1: Power curves based on the number of replicates per condition for the six public datasets and five RNA-Seq differential expression analysis packages. Genes were further subcategorized by gene expression and fold change (FC) levels. All genes: the proportion of detected DE genes. High vs. low FC: upper vs. lower quartile based on FC of DE genes. High vs. low expression: upper vs. lower quartile based on expression level of DE genes. Library sizes were estimated from the gene counts of the real datasets. FC and expression levels were estimated through a general linear model with negative binomial distribution. Per-gene dispersion was estimated through the Cox-Reid adjusted profile likelihood. The four unpaired-sample datasets (Bottomly, Bullard, Huang, M-P) were analyzed with edgeR, DESeq, DESeq2, EBSeq and sSeq. The paired-sample datasets (Tuch and Qian) were analyzed with edgeR, DESeq, DESeq2 and sSeq. Note that EBSeq is not included as it is currently not adapted to analyzing paired-sample data.

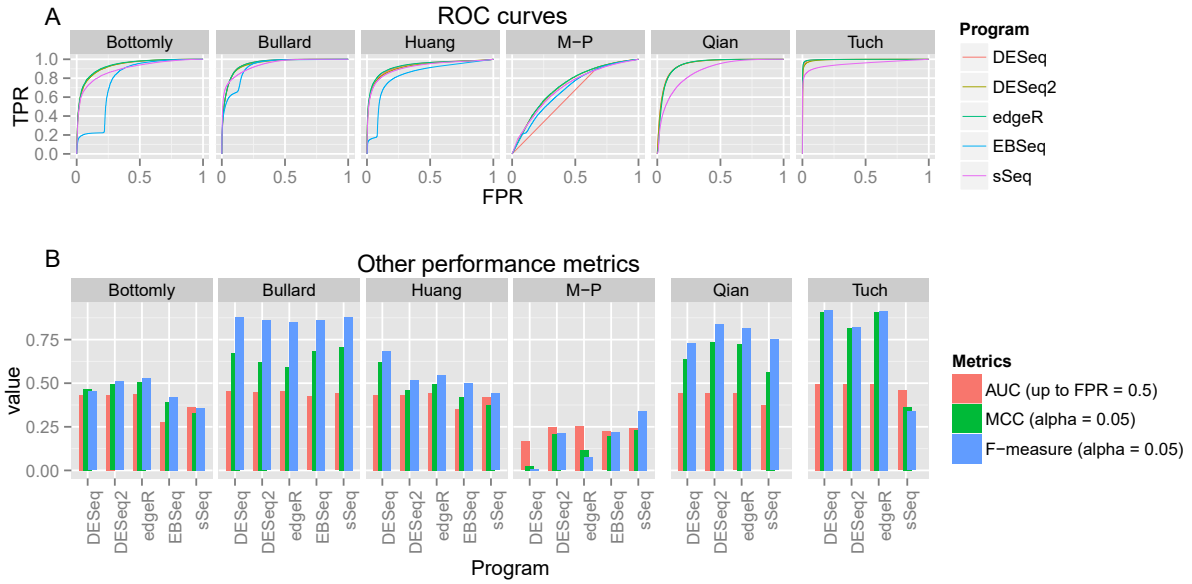


Figure 2: Performance comparison with Receiver Operator Characteristics (ROC) curves and other metrics for the six public datasets and five RNA-Seq differential expression analysis packages. Sensitivity and 1-specificity were estimated in each simulation for n=4 replicates per condition. The simulations were conducted as in Fig. 1. A. ROC curve comparison. True Positive Rate (TPR) vs. False Positive Rate (FPR) was plotted. B. Other performance metrics. Area under the curve (AUC) was measured up to FPR = 0.5 of the ROC curves in A. Matthew Correlation Coefficient (MCC) and F-measure were measured at the threshold of alpha = 0.05.

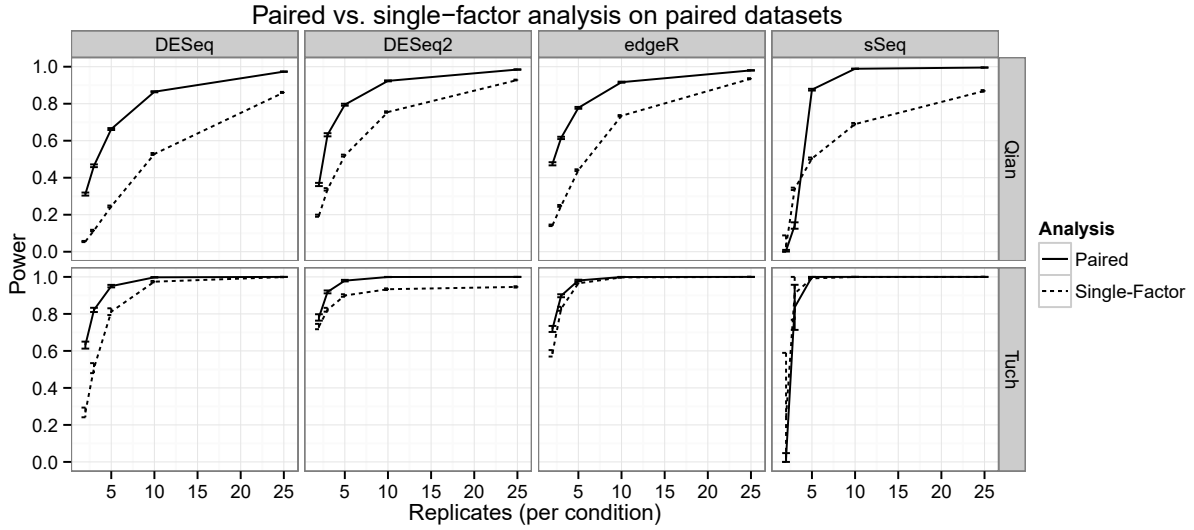


Figure 3: Paired vs. single-factor power analysis of paired-sample datasets (Qian and Tuch). The datasets were evaluated with pairing information (paired analysis, solid line) or without pairing information (single-factor analysis, dashed line), using the standard analysis pipelines for the respective packages as in Fig. 1. Note that EBSeq is not included as it is currently not adapted to analyzing paired-sample data.

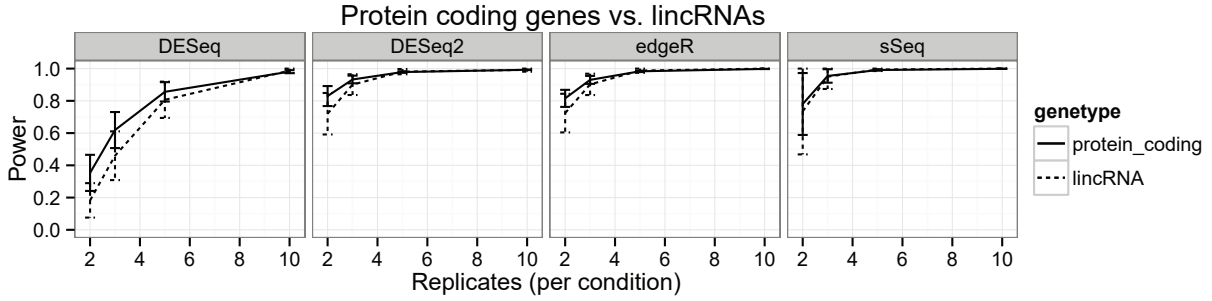


Fig. 4: Power of protein coding genes vs. long non-coding RNA (lincRNA) transcripts. The comparison was made using the Huang dataset, which used ribosomal RNA removal for RNA library construction. The transcriptome was separated into protein coding genes (solid line) or lincRNA (dashed line) categories. Power was estimated in each simulation for these two categories, using the standard analysis pipelines for the respective packages as in Fig. 1.

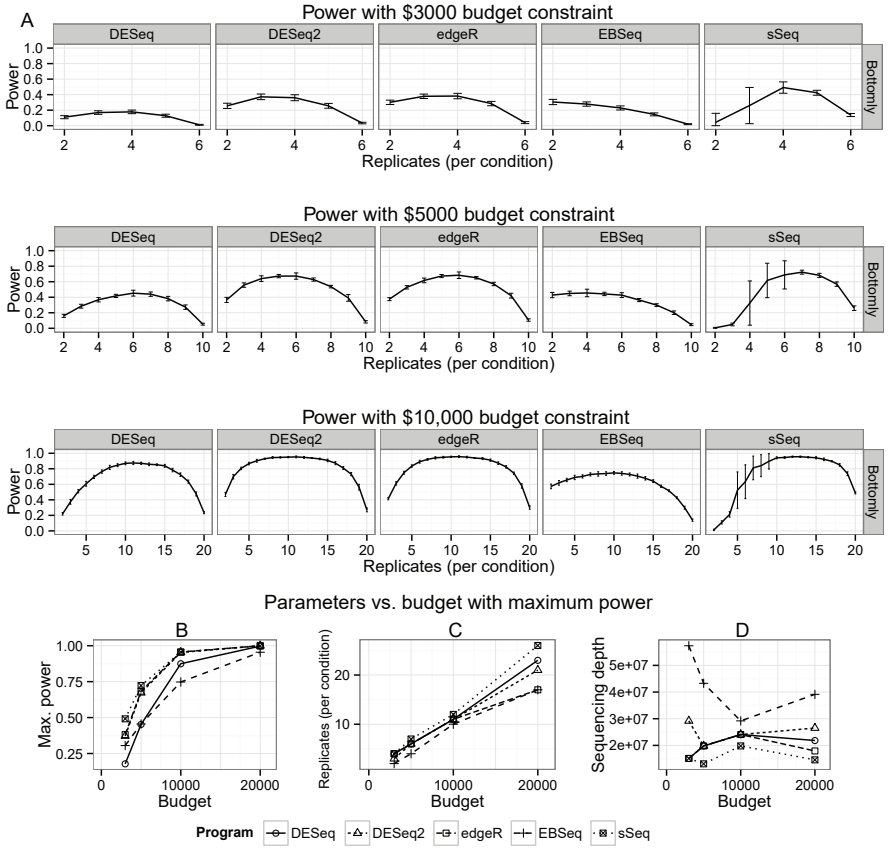


Fig. 5: Optimization of power given a budget constraint. The cost of RNA-Seq per sample is dependent on the cost of constructing the RNA-Seq library and the cost of single-end sequencing under the multiplex arrangement, where multiple samples could be barcoded to share one lane of the HiSeq flow cell. Both sequencing depth and sample size are variables under the budget constraint. A: power curves relative to replicates, exemplified by increasing budgets of $3000, $5000 and $10000 among five RNA-Seq differential expression analysis packages. B: Optimal powers achieved for given budget constraints. C: Biological replicates required to obtain optimal powers for given budget constraints. D: Sequencing depths required to obtain optimal powers for given budget constraints.

Table 1: Description of the six public RNA-Seq datasets and estimation of dataset parameters

Dataset	RNA selection	Description	Experimental Design	Organism	Number of samples	median expression (log2 counts per million)
Bottomly	Poly-A enrichment	Expression comparison between two cell lines	one factor	Mouse	21	3.96 (1.66 - 6.16)
Bullard	Poly-A enrichment	Comparison of brain tissue to reference RNA library (2 biological samples, 7 technical replicates)	one factor	Human	14	4.72 (2.72 - 6.65)
Huang	Ribozyme depletion	Comparison of CCE cells vs. Fetal Head cells (4 biological samples)	one factor	Mouse	22	4.37 (1.54 - 6.1)
M-P	Poly-A enrichment	Comparison between Caucasian and Nigerian populations	one factor	Human	129	4.05 (1.08 - 6.61)
Qian	Ribozyme depletion	WNV/Mock transfection comparison	paired samples	Human	20	4.3 (1.57 - 6.14)
Tuch	Poly-A enrichment	Cancer/Normal Tissue comparison	paired samples	Human	6	5.22 (4.11 - 6.3)

Dataset	median log2 fold change of DE genes	median dispersion	Percent DE	Mean Library Size (sum of total counts, log 10)	Average sequencing depth (log10)	Percent map to transcriptome (after filtering)
Bottomly	0.99 (0.615 - 1.69)	0.035 (0.0153 - 0.0756)	4.29%	6.67 +/- 0.02	7.34	21.3%
Bullard	2.13 (1.28 - 3.85)	0.000391 (0.000391 - 0.00488)	59.30%	6.11 +/- 0.00134	7.10	10.3%
Huang	3.33 (2.28 - 4.67)	0.128 (0.0594 - 0.25)	15.64%	6.16 +/- 0.143	7.25	8.2%
M-P	0.751 (0.575 - 1.1)	0.231 (0.11 - 0.724)	21.50%	6.17 +/- 0.0341	7.23	8.8%
Qian	0.929 (0.627 - 1.51)	0.0454 (0.0167 - 0.0959)	44.23%	7.00 +/- 0.00312	7.45	35.7%
Tuch	2.13 (1.68 - 2.94)	0.0776 (0.0285 - 0.173)	7.97%	6.97 +/- 0.031	8.31	3.4%

Table 2: DE genes (FDR ¡ 0.05) detected by the different analysis packages

Dataset	Total Genes	DESeq2	DESeq	edgeR	sSeq	EBSeq	Intersection	Percent DE
Bottomly	10645	1348	588	1221	1200	579	457	4.29%
Bullard	9100	7573	7381	7667	6371	5973	5396	59.30%
Huang	17872	9842	3306	10062	12291	8308	2795	15.64%
Montgomery-Pickrell	9217	5014	2964	5264	3553	3018	1982	21.50%
Qian	17110	9670	8098	9404	16442	N/A	7567	44.23%
Tuch	15668	2072	1340	1903	5011	N/A	1248	7.97%

Table 3: Estimated parameters of protein-coding genes vs. lincRNA transcripts

	Total Number	Differentially Expressed	median gene expression (log2 counts per million + 1)	median log2 fold change of DE genes	median dispersion	percent DE
All genes	17872	2795	4.37 (1.54 - 6.1)	3.33 (2.28 - 4.67)	0.128 (0.0594 - 0.25)	0.15638988
Protein coding	15834	2623	4.63 (2.05 - 6.23)	3.34 (2.27 - 4.67)	0.126 (0.0599 - 0.242)	0.16565618
lincRNA	603	79	1.18 (-1.43 - 2.71)	3.32 (2.57 - 4.58)	0.139 (0.0406 - 0.285)	0.13101161

3.7 Appendix

3.7.1 Supplementary figures and tables

Figure S1: Power curves based on the sequencing depths for the six public datasets and different RNA-Seq differential expression analysis packages. Power was compared in 10, 5 and 3 replicates respetively. The four unpaired datasets (Bottomly, Bullard, Huang, and M-P) were analyzed with edgeR, DESeq, DESeq2, EBSeq and sSeq. The paired datasets (Tuch and Qian) were analyzed with edgeR, DESeq, DESeq2 and sSeq. Note that EBSeq is not included as it is currently not adapted to analyzing paired-sample data.

Figure S2: Extended performance comparison with Receiver Operator Characteristics (ROC) curves for the five categories of DE genes from six public datasets using five RNA-Seq differential expression analysis packages. True Positive Rate (TPR) vs. False Positive Rate (FPR) was plotted. All genes: the proportion of detected DE genes. High vs. low FC: upper vs. lower quartile based on FC of DE genes. High vs. low expression: upper vs. lower quartile based on expression level of DE genes.

Figure S3: Extended performance comparison with other metrics (AUC, MCC and F-measure) for the five categories of DE genes from six public datasets using five RNA-Seq differential expression analysis packages. Area under the curve (AUC) was measured up to FPR = 0.5 of the ROC curves in Fig. S2. Matthew Correlation Coefficient (MCC) and F-measure were measured at the threshold of alpha = 0.05.

Figure S4: Power curves relative to replicates, given budget constraints for all six datasets. Power curves are demonstrated by increasing budgets of \$3000, \$5000, \$10000 and \$20,000, estimated by five RNA-Seq differential expression analysis packages.

Table S1: the AUC, MCC and F-measure values for the six public datasets and five RNA-Seq differential expression analysis packages.

Figure S1

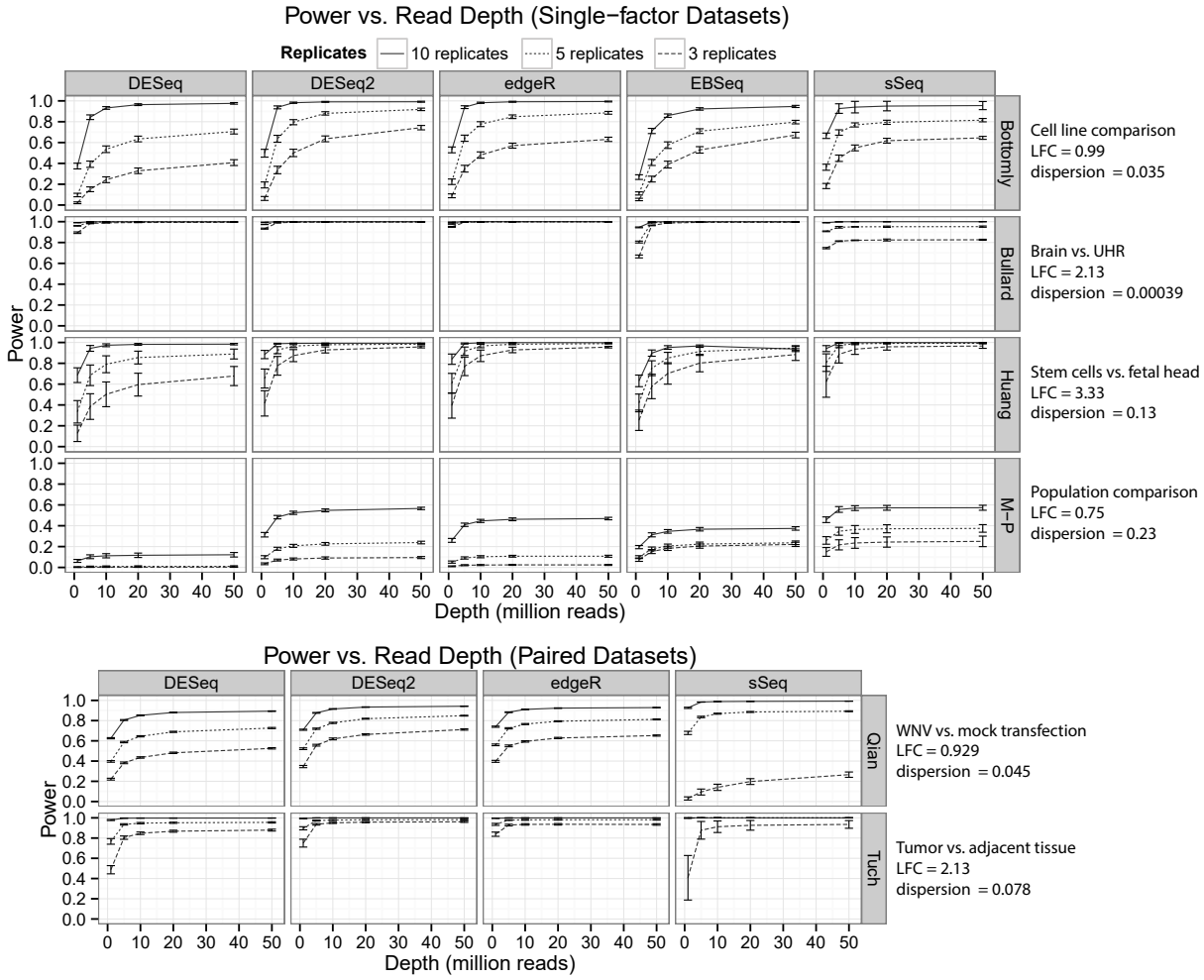

Power vs. Read Depth (Single−factor Datasets)
Replicates
10 replicates
5 replicates
3 replicates
DESeq
DESeq2
edgeR
EBSeq
sSeq
Bottomly
Bullard
Huang
M−P
Power
Depth (million reads)
Cell line comparison
LFC = 0.99
dispersion = 0.035
Brain vs. UHR
LFC = 2.13
dispersion = 0.00039
Stem cells vs. fetal head
LFC = 3.33
dispersion = 0.13
Population comparison
LFC = 0.75
dispersion = 0.23
Power vs. Read Depth (Paired Datasets)
Qian
Tuch
WNV vs. mock transfection
LFC = 0.929
dispersion = 0.045
Tumor vs. adjacent tissue
LFC = 2.13
dispersion = 0.078

Figure S2

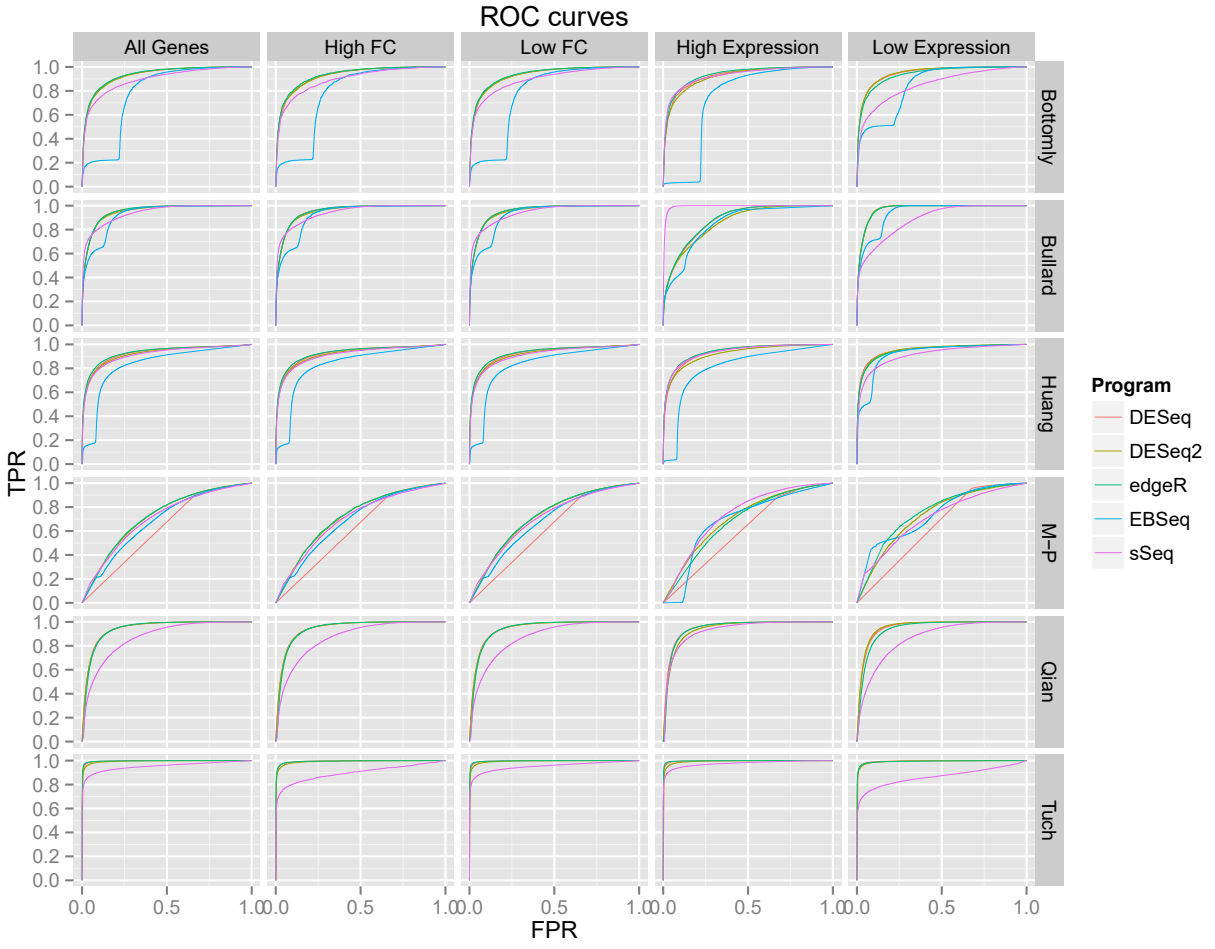

ROC curves
All Genes
High FC
Low FC
High Expression
Low Expression
Bottomly
Bullard
Huang
M-P
Qian
Tuch
TPR
FPR
0.0
0.2
0.4
0.6
0.8
1.0
0.5
Program
DESeq
DESeq2
edgeR
EBSeq
sSeq

Figure S3

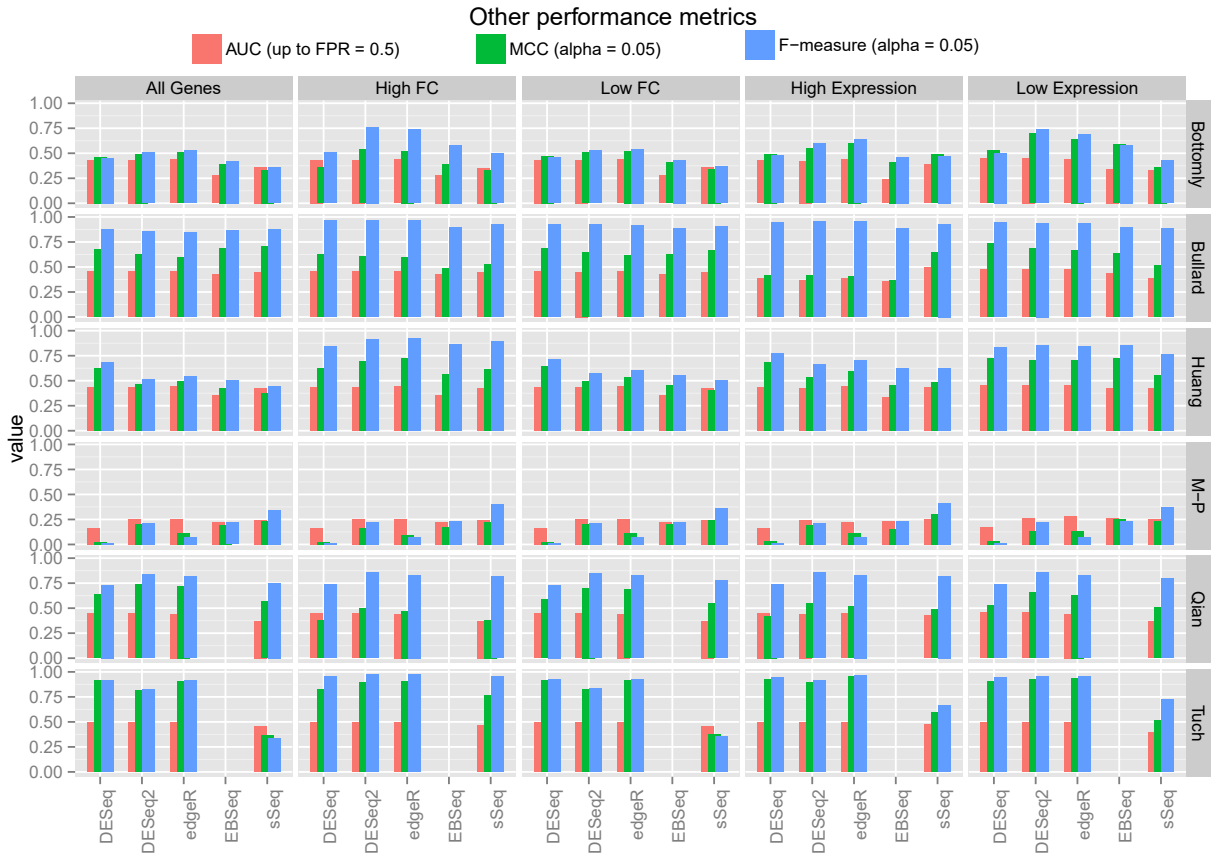

Other performance metrics
AUC (up to FPR = 0.5)
MCC (alpha = 0.05)
F−measure (alpha = 0.05)
All Genes
High FC
Low FC
High Expression
Low Expression
Bottomly
Bullard
Huang
M−P
Qian
Tuch
value
1.00
0.75
0.50
0.25
0.00
DESeq
DESeq2
edgeR
EBSeq
sSeq

Figure S4.1

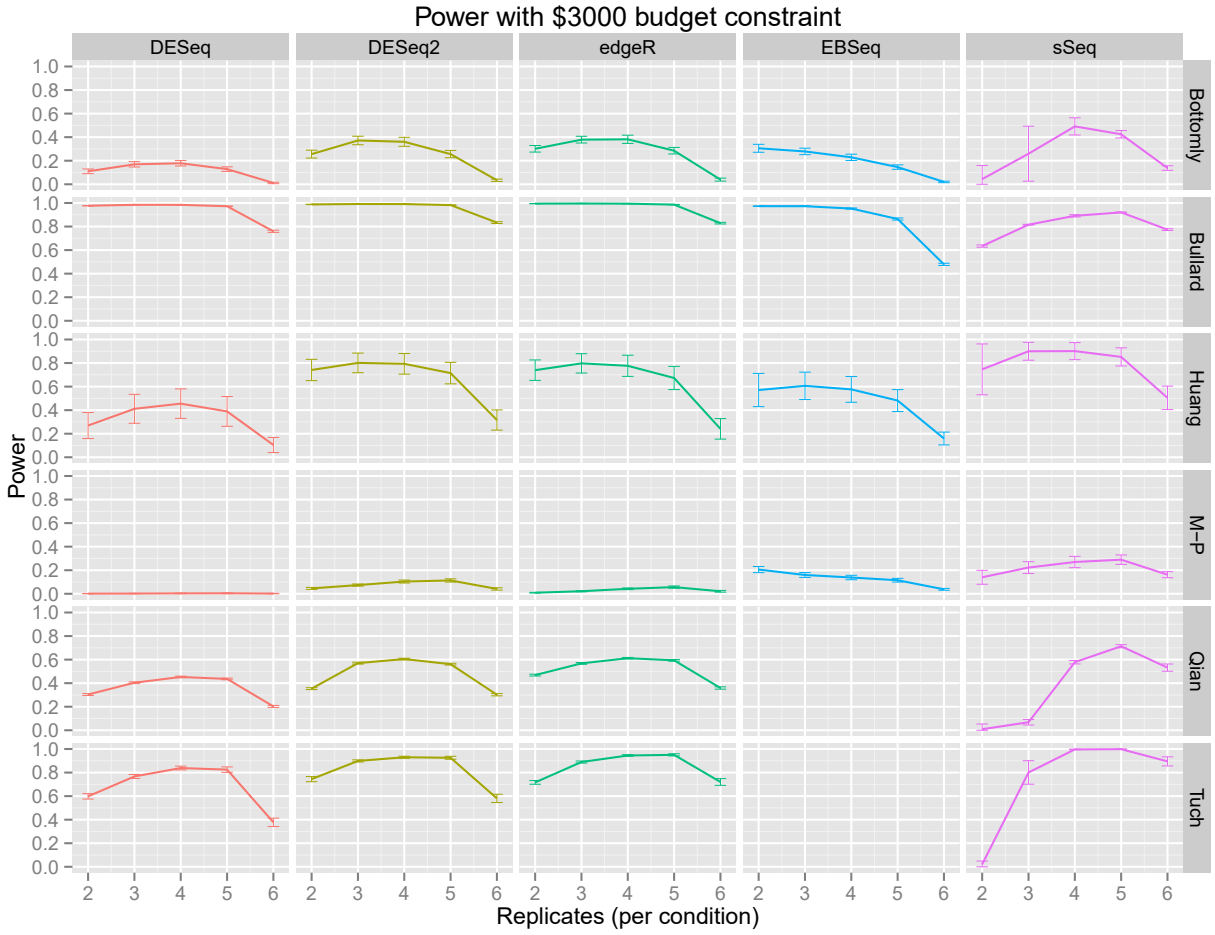

Power with $3000 budget constraint
DESeq
DESeq2
edgeR
EBSeq
sSeq
Bottomly
Bullard
Huang
M−P
Qian
Tuch
Power
Replicates (per condition)

Figure S4.2

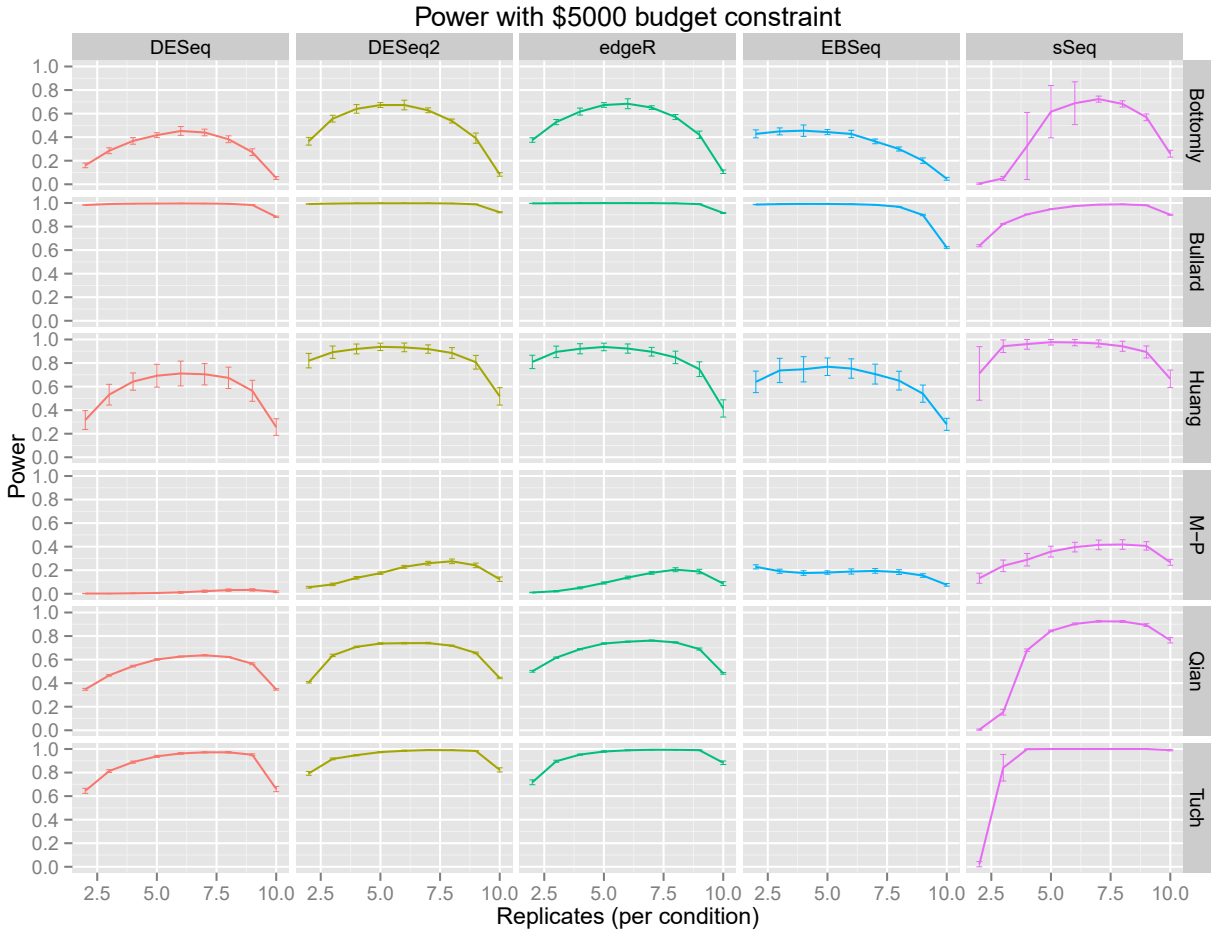

Power with $5000 budget constraint
DESeq
DESeq2
edgeR
EBSeq
sSeq
Bottomly
Bullard
Huang
M−P
Qian
Tuch
Power
Replicates (per condition)
0.0
0.2
0.4
0.6
0.8
1.0
2.5
5.0
7.5
10.0

Figure S4.3

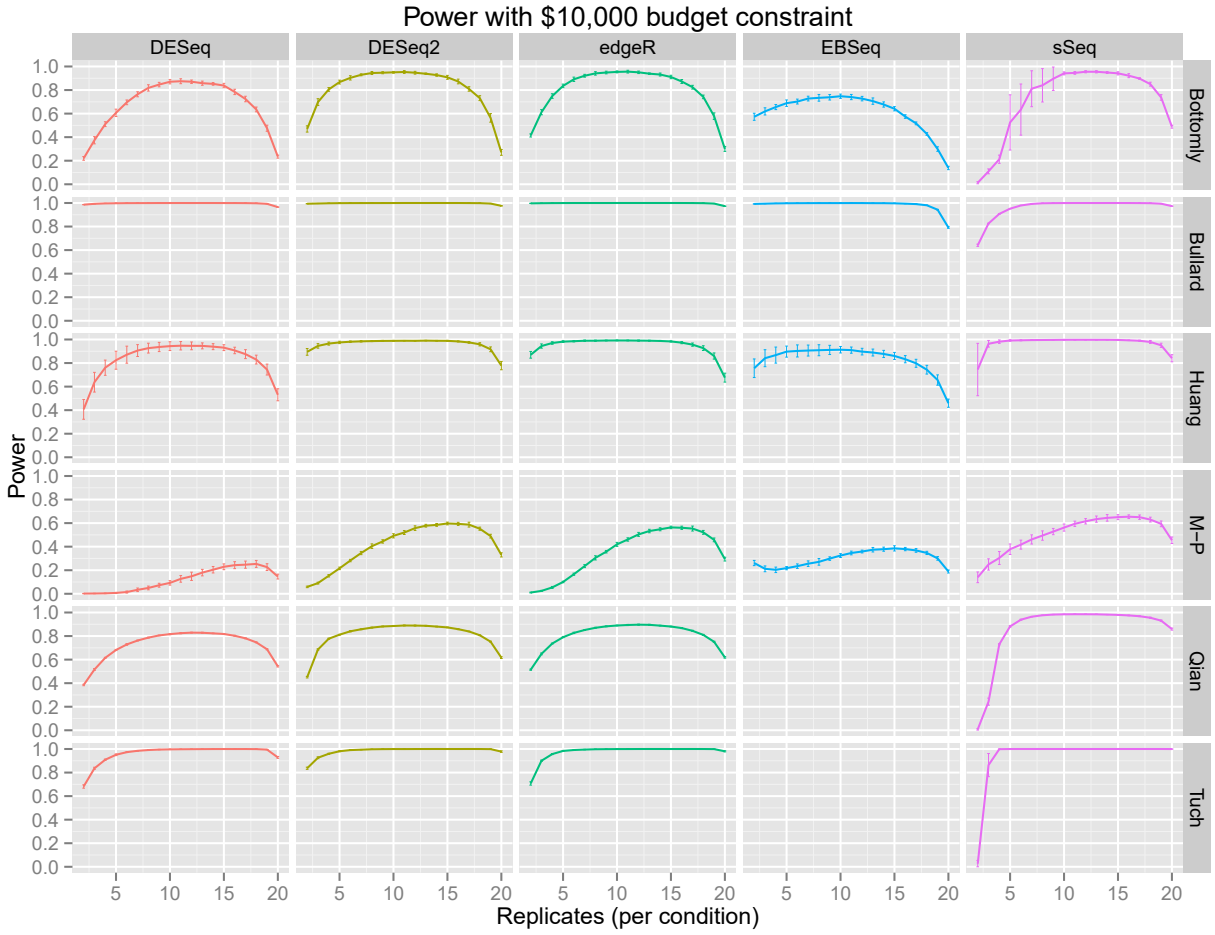

Power with $10,000 budget constraint
DESeq
DESeq2
edgeR
EBSeq
sSeq
Bottomly
Bullard
Huang
M−P
Qian
Tuch
Power
Replicates (per condition)
0.0
0.2
0.4
0.6
0.8
1.0
5
10
15
20

Table S1.1

Dataset	Program	Subset	AUC (up to FPR = 0.5)	MCC (alpha = 0.05)	F-measure (alpha = 0.05)
Bottomly	DESeq2	All Genes	0.433285586	0.492072291	0.509437612
Bottomly	DESeq2	High FC	0.432672461	0.539585716	0.75993514
Bottomly	DESeq2	Low FC	0.433343466	0.507926519	0.529210806
Bottomly	DESeq2	High Expression	0.421926238	0.549428279	0.595823996
Bottomly	DESeq2	Low Expression	0.450197884	0.699839292	0.736084121
Bottomly	DESeq	All Genes	0.433018354	0.464943779	0.452827987
Bottomly	DESeq	High FC	0.433224208	0.358488855	0.505883503
Bottomly	DESeq	Low FC	0.43305405	0.472623674	0.458263206
Bottomly	DESeq	High Expression	0.427292061	0.493080913	0.477964953
Bottomly	DESeq	Low Expression	0.448442893	0.527663428	0.501146308
Bottomly	edgeR	All Genes	0.436827938	0.506603199	0.526849195
Bottomly	edgeR	High FC	0.436528332	0.525229544	0.73854712
Bottomly	edgeR	Low FC	0.436896646	0.521751225	0.544730491
Bottomly	edgeR	High Expression	0.435917098	0.601260875	0.637098197
Bottomly	edgeR	Low Expression	0.439051199	0.642226971	0.686862027
Bottomly	sSeq	All Genes	0.363107949	0.330670032	0.356730264
Bottomly	sSeq	High FC	0.353354491	0.331423003	0.497065088
Bottomly	sSeq	Low FC	0.364012207	0.342940416	0.369225195
Bottomly	sSeq	High Expression	0.388655872	0.489209554	0.472868457
Bottomly	sSeq	Low Expression	0.330667672	0.360105406	0.427199933
Bottomly	EBSeq	All Genes	0.278989459	0.394290591	0.420182659
Bottomly	EBSeq	High FC	0.280569174	0.393898242	0.585249365
Bottomly	EBSeq	Low FC	0.278997527	0.406843497	0.43392693
Bottomly	EBSeq	High Expression	0.236929017	0.410739265	0.462168925
Bottomly	EBSeq	Low Expression	0.344721108	0.588723372	0.58121893
Bullard	DESeq2	All Genes	0.450341286	0.620665066	0.859118012
Bullard	DESeq2	High FC	0.45123677	0.60783282	0.965112805
Bullard	DESeq2	Low FC	0.450106257	0.640893014	0.92134452
Bullard	DESeq2	High Expression	0.364495656	0.412906325	0.950322021
Bullard	DESeq2	Low Expression	0.471612474	0.684164971	0.939992895
Bullard	DESeq	All Genes	0.454003092	0.673753018	0.87700467
Bullard	DESeq	High FC	0.454728643	0.623272766	0.962878883
Bullard	DESeq	Low FC	0.453817133	0.68114862	0.927751429
Bullard	DESeq	High Expression	0.381739538	0.41331528	0.946373396
Bullard	DESeq	Low Expression	0.471638353	0.733697942	0.946729776
Bullard	edgeR	All Genes	0.454098772	0.591592976	0.849528091
Bullard	edgeR	High FC	0.454982887	0.594981406	0.965372338
Bullard	edgeR	Low FC	0.453875701	0.617177574	0.917326981
Bullard	edgeR	High Expression	0.383010239	0.401253813	0.951084936
Bullard	edgeR	Low Expression	0.471563072	0.667083276	0.937664847
Bullard	sSeq	All Genes	0.441209514	0.70548017	0.878297797
Bullard	sSeq	High FC	0.441344849	0.525771882	0.921195625
Bullard	sSeq	Low FC	0.441687067	0.664368423	0.904548261
Bullard	sSeq	High Expression	0.49278953	0.640243022	0.929741084
Bullard	sSeq	Low Expression	0.387562656	0.514681108	0.888396762
Bullard	EBSeq	All Genes	0.423757506	0.685959677	0.862252614
Bullard	EBSeq	High FC	0.424343874	0.482043907	0.8970692
Bullard	EBSeq	Low FC	0.42344753	0.629520633	0.883233346
Bullard	EBSeq	High Expression	0.356588323	0.363562391	0.885268112
Bullard	EBSeq	Low Expression	0.439068102	0.639188789	0.892477647
Huang	DESeq2	All Genes	0.432693774	0.4600306	0.516492629
Huang	DESeq2	High FC	0.433360662	0.691315035	0.916415628

Table S1.2

Huang	DESeq2	Low FC	0.432704065	0.495532439	0.575773073
Huang	DESeq2	High Expression	0.427995157	0.536140975	0.661774779
Huang	DESeq2	Low Expression	0.458054312	0.706752199	0.853507246
Huang	DESeq	All Genes	0.432590661	0.62154443	0.685277416
Huang	DESeq	High FC	0.433419322	0.621192688	0.844780359
Huang	DESeq	Low FC	0.432586626	0.644332982	0.717377624
Huang	DESeq	High Expression	0.433476842	0.683047814	0.770911087
Huang	DESeq	Low Expression	0.45356313	0.723982986	0.831414237
Huang	edgeR	All Genes	0.442319802	0.49663027	0.546295247
Huang	edgeR	High FC	0.443389363	0.721115613	0.921528845
Huang	edgeR	Low FC	0.442258301	0.532975541	0.604370763
Huang	edgeR	High Expression	0.445031728	0.59011258	0.701076832
Huang	edgeR	Low Expression	0.453226345	0.704131542	0.841659537
Huang	sSeq	All Genes	0.420638113	0.375267652	0.441088419
Huang	sSeq	High FC	0.420523618	0.61754034	0.897222985
Huang	sSeq	Low FC	0.420561831	0.407533294	0.501061211
Huang	sSeq	High Expression	0.432390622	0.485172724	0.625693779
Huang	sSeq	Low Expression	0.421399021	0.553533932	0.765882962
Huang	EBSeq	All Genes	0.353430082	0.422430661	0.501939494
Huang	EBSeq	High FC	0.353407647	0.562501785	0.868726144
Huang	EBSeq	Low FC	0.3534716	0.453516988	0.557669298
Huang	EBSeq	High Expression	0.330557595	0.458555023	0.624160627
Huang	EBSeq	Low Expression	0.423916185	0.724793125	0.85198609
M-P	DESeq2	All Genes	0.251015458	0.208860567	0.211581765
M-P	DESeq2	High FC	0.252003023	0.16071344	0.221112221
M-P	DESeq2	Low FC	0.250841952	0.207119946	0.215472482
M-P	DESeq2	High Expression	0.242131288	0.19718241	0.218327689
M-P	DESeq2	Low Expression	0.26139684	0.130610037	0.221911536
M-P	DESeq	All Genes	0.165780296	0.027414877	0.006561726
M-P	DESeq	High FC	0.16607886	0.025467023	0.006578483
M-P	DESeq	Low FC	0.165595508	0.027715166	0.006568194
M-P	DESeq	High Expression	0.162036359	0.031145416	0.006578285
M-P	DESeq	Low Expression	0.172543932	0.038447598	0.006578384
M-P	edgeR	All Genes	0.255088872	0.114953176	0.07574754
M-P	edgeR	High FC	0.255954002	0.095571342	0.077327306
M-P	edgeR	Low FC	0.254840444	0.116588966	0.076406918
M-P	edgeR	High Expression	0.228334476	0.110183172	0.077261255
M-P	edgeR	Low Expression	0.281337751	0.131511948	0.077126519
M-P	sSeq	All Genes	0.244551405	0.230438039	0.340876536
M-P	sSeq	High FC	0.245193945	0.221263833	0.407081818
M-P	sSeq	Low FC	0.244467183	0.244204886	0.366639837
M-P	sSeq	High Expression	0.257695521	0.300111673	0.41040791
M-P	sSeq	Low Expression	0.251448588	0.233443921	0.378137459
M-P	EBSeq	All Genes	0.223569194	0.196150379	0.219533557
M-P	EBSeq	High FC	0.223935579	0.172279563	0.236457136
M-P	EBSeq	Low FC	0.22329616	0.201461103	0.226370552
M-P	EBSeq	High Expression	0.236394182	0.151892643	0.231735111
M-P	EBSeq	Low Expression	0.263541891	0.253033705	0.236150428
Qian	DESeq2	All Genes	0.444051105	0.737859031	0.838512035
Qian	DESeq2	High FC	0.443740005	0.494005715	0.855675801
Qian	DESeq2	Low FC	0.444170163	0.702946998	0.846409898
Qian	DESeq2	High Expression	0.441003056	0.543703994	0.854315978
Qian	DESeq2	Low Expression	0.45793853	0.658687765	0.855743089
Qian	DESeq	All Genes	0.444985522	0.637771086	0.730564227

Table S1.3

Qian	DESeq	High FC	0.444629608	0.376091854	0.735206263
Qian	DESeq	Low FC	0.445130108	0.5889196	0.732705396
Qian	DESeq	High Expression	0.453206186	0.422273459	0.735227273
Qian	DESeq	Low Expression	0.453913422	0.528529092	0.735229127
Qian	edgeR	All Genes	0.441164028	0.723253842	0.818593973
Qian	edgeR	High FC	0.440845566	0.470537391	0.830404718
Qian	edgeR	Low FC	0.441251095	0.683795474	0.824004371
Qian	edgeR	High Expression	0.448369579	0.518472209	0.830317115
Qian	edgeR	Low Expression	0.440657485	0.631465766	0.827682341
Qian	sSeq	All Genes	0.372126128	0.566493446	0.751034542
Qian	sSeq	High FC	0.371833136	0.382458144	0.816156676
Qian	sSeq	Low FC	0.37203309	0.548945601	0.779590543
Qian	sSeq	High Expression	0.428449434	0.485557003	0.821261729
Qian	sSeq	Low Expression	0.366759882	0.506482001	0.796306927
Tuch	DESeq2	All Genes	0.493792209	0.815019767	0.823454026
Tuch	DESeq2	High FC	0.49356017	0.889981969	0.973280236
Tuch	DESeq2	Low FC	0.49378378	0.824926127	0.835003755
Tuch	DESeq2	High Expression	0.49333735	0.896626091	0.91905443
Tuch	DESeq2	Low Expression	0.493581205	0.923279101	0.955114143
Tuch	DESeq	All Genes	0.494273431	0.910104034	0.917351558
Tuch	DESeq	High FC	0.49301075	0.825494182	0.950691461
Tuch	DESeq	Low FC	0.494291613	0.912625422	0.920227758
Tuch	DESeq	High Expression	0.494625502	0.927928968	0.942156802
Tuch	DESeq	Low Expression	0.492590769	0.908701127	0.943808992
Tuch	edgeR	All Genes	0.496622751	0.907957823	0.914823411
Tuch	edgeR	High FC	0.496391957	0.907276573	0.976998305
Tuch	edgeR	Low FC	0.496631139	0.9128054	0.92009337
Tuch	edgeR	High Expression	0.497462851	0.954263613	0.964124675
Tuch	edgeR	Low Expression	0.493046637	0.930734649	0.959381215
Tuch	sSeq	All Genes	0.458072739	0.36269168	0.337614694
Tuch	sSeq	High FC	0.460482487	0.768541612	0.95163797
Tuch	sSeq	Low FC	0.457828518	0.375211241	0.357664747
Tuch	sSeq	High Expression	0.47550779	0.598056121	0.66456535
Tuch	sSeq	Low Expression	0.396962457	0.515187054	0.725999942

3.8 Chapter summary

In this chapter, we compare different statistical methods for finding differentially expressed genes. In finding differentially expressed genes, generally, one must consider two measures: statistical power (i.e., true positive rate) with false discoveries (i.e., non-differential genes that are incorrectly detected). These two measures can be balanced through summary statistics, such as area under the reciever operator curve (AUC), Matthew's correlation coefficient and F1-measure. As a result, we prove that two methods stand above other competitors. These two methods are implemented in software packages DESeq2 and edgeR.

Understanding of RNA-Seq and its downstream analysis is essential for studying lincRNAs from high throughput sequencing. We use these packages as a means of obtaining normalized lincRNA expression between samples and finding differentially expressed lincRNAs in order to achieve the specific aims.

Chapter 4

Pan-cancer analyses reveal lincRNAs relevant to tumour diagnosis, subtyping and prognosis

Travers Ching[1,2], Karolina Peplowska[3], Sijia Huang[1,2], Xun Zhu[1,2], Yi Shen[2], Janos Molnar[3], Herbert Yu[2], Maarit Tiirikainen[3], Ben Fogelgren[4], Rong Fan[5], Lana X Garmire[1,2]

Published in *EBioMedicine* (2016).

[1]University of Hawaii Cancer Center, 701 Ilalo Street, Honolulu, Hawaii, USA 96813

[2]Graduate Program of Molecular Biosciences and Bioengineering, University of Hawaii-Manoa, 1955 East-West Road, Honolulu, Hawaii, USA 96822

[3]Genomics Shared Resource, University of Hawaii Cancer Center, Honolulu, HI, 96813, USA

[4]Department of Anatomy, Biochemistry and Physiology, John A. Burns School of Medicine, University of Hawaii, Honolulu, HI 96813, USA

[5]Department of Biomedical Engineering, Yale University, New Haven, CT 06520, USA

4.1 Preface

Long intergenic noncoding RNAs (lincRNAs) are a relatively new class of non-coding RNAs that have the potential as cancer biomarkers. However, currently most lincRNA biomarkers are

detected in individual cancer types, and their pan-cancer biomarker application has not yet been reported. To seek a panel of lincRNAs as pan-cancer biomarkers, we have analyzed transcriptomes from over 3300 cancer samples with clinical information. Compared to mRNA, lincRNAs exhibit significantly higher tissue specificities that are then diminished in cancer tissues. Moreover, lincRNA clustering results accurately classify tumour subtypes. Using RNA-Seq data from thousands of paired tumour and adjacent normal samples in The Cancer Genome Atlas (TCGA), we identify six lincRNAs as potential pan-cancer diagnostic biomarkers (PCAN-1, PCAN-2, PCAN-3, PCAN-4, PCAN-5 and PCAN-6). These lincRNAs are robustly validated using cancer samples from four independent RNA-Seq data sets, and are verified by qPCR in both primary breast cancers and MCF-7 cell line. Interestingly, the expression levels of these six lincRNAs are also associated with prognosis in various cancers. We performed cell line experiments on two breast cancer cell lines suggesting that growth and migration are affected by the expression of these lincRNAs. In summary, our study highlights the emerging role of lincRNAs as potentially powerful and biologically functional pan-cancer biomarkers and represents a significant leap forward in understanding the biological and clinical functions of lincRNAs in cancers.

4.2 Introduction

Advancement of high-throughput technologies such as RNA-Seq has recently allowed for the identification of tens of thousands of new lincRNAs in different tissues [1–4]. The Encyclopedia of DNA Elements (ENCODE) project found that about 62% of the entire genome is transcribed to long ($>$ 200 base pairs) RNA sequences [5]. Given that 3% of the genome encodes protein-coding exons, the large majority of these transcripts are non-coding RNAs (lncRNAs). Among these lncRNAs, about one third come from intergenic regions (lincRNAs) [5]. Unlike small non-coding RNAs which may regulate target gene expression through simpler complementary recognition [6], the mechanisms of lincRNAs are complex and may depend on formation of RNA-protein complexes [7]. Attempts have been made to extrapolate the functions of lincRNAs based on model lincRNAs, such as studies that predict lincRNAs binding to PRC2 or competing endogenous lincRNAs (micro-RNA "sponges") [8–12]. However, lincRNAs remain one of the most mysterious and least understood species of non-coding RNAs [2].

Regardless of the regulatory mechanisms, lincRNAs are becoming a relatively new class of cancer biomarker candidates. Several lincRNAs and overlapping lncRNAs have been relatively well-studied and indicated as potential biomarkers associated with tumour initiation, progression or prognosis, such as MALAT [13–15], HOTAI [15–17], XIS [18–20], PCAT [15, 21, 22] and CCAT

[23]. However, most of the studies detect lincRNAs as candidate biomarkers of a specific cancer type. The pan-cancer biomarker-based design of clinical trials, on the other hand, can increase statistical power and greatly decreasing the size, expense, and duration of clinical trials [24]. Towards this, we here propose a pan-cancer based lincRNA diagnostics biomarker study, which is aligned with the goal of TCGA analysis project that enables the discovery of novel adaptive, biomarker-based strategies to be practiced across boundaries of different tumour type [24] . In this study, we have taken full advantage of the rich RNA-Seq data from the TCGA consortium, thousands of RNA-Seq and microarray data from Gene Expression Omnibus (GEO) as well as RNA-Seq data from our own collection of breast cancer samples. By combining data-mining and machine-learning methods with biological function validation experiments, we have highlighted lincRNAs as a new paradigm for actionable diagnostics in the pan-cancer setting. In addition, we have portrayed the comprehensive landscape of lincRNAs and their relationship to other omics data in pan-cancers. We found that the lincRNAs are more tissue-specific compared to protein-coding mRNAs, and they also convey complementary relevance to clinical information, including tumour molecular subtypes. Moreover, we have detected and thoroughly validated 6 lincRNAs as potential pan-cancer diagnostic biomarkers in over 3300 tissue samples. Finally, we confirmed that the lincRNAs are biologically functional, by measuring the reduction of cell proliferation and migration in breast cancer cell lines with siRNA knockdown on two of the homologous lincRNAs.

4.3 Methods

4.3.1 RNA-Seq datasets

TCGA datasets

We used 12 cancer datasets from TCGA incorporating RNA-Seq data files from 1240 tissue samples (Supplementary Table I). RNA-Seq datasets were chosen from cancers in TCGA that have at least 25 pairs of primary tumour and paired adjacent normal tissue samples. These datasets include breast invasive carcinoma (BRCA), colon adenocarcinoma (COAD), head and neck squamous cell carcinoma (HSNC), kidney chromophobe (KICH), kidney renal clear cell carcinoma (KIRC), kidney renal papillary cell carcinoma (KIRP), lung adenocarcinoma (LUAD), lung squamous cell carcinoma (LUSC), prostate adenocarcinoma (PRAD), stomach adenocarcinoma (STAD) and thyroid carcinoma (THCA). RNA-Seq BAM files were downloaded from UCSC Cancer Genomics Hub (`https://cghub.ucsc.edu/`) using the GeneTorrent progra [25]. The TCGA alignment protocol used the Mapsplice alignment progra [26] to align raw reads to

the human genome, where loci with the same alignment score has equal probability to assign a read. Technical replicates were combined by merging the results from the BAM files. RefSeq genes and lincRNAs were quantified using featureCounts [27, 28] from the Subread package (version 1.4.5-p1). RefSeq annotation was obtained from Illumina hg19 iGenomes and lincRNAs were obtained from Broad Institute Human Body Map project, so that we can directly compare the tissue specificity results between TCGA samples and those in Cabili et al [1]. All alignments were conducted on the New Hampshire INBRE (IDeA Network of Biomedical Research Excellence) grid computing system. DESeq [29, 30] (version 1.6.1) was used for calculating normalized count data and FPKM data. A combination of independent RNA-Seq and microarray datasets were used for verification, and the summary of the datasets is listed in Supplementary Table I.

GEO datasets

A comprehensive search of GEO RNA-Seq database was performed to find additional datasets for verification. Datasets with tumour and normal samples with good read quality (read mapping rate and low duplication rates) were selected. These included GSE25599 (liver cancer), GSE58135 (breast cancer) and GSE50760 (colon cancer). In addition, normal breast tissue samples were taken from GSE52194, GSE45326 and GSE30611 for comparison with our cancer samples. GEO datasets were aligned to the UCSC hg19 genome using Tophat2 with default parameters for either single-end or paired-end protocols. LincRNA count quantification and FPKM data were generated as above. Microarray datasets from GEO with tumour and normal samples were selected based on platforms that had probes mapping to the six lincRNAs of interest.

Our own dataset

Our primary breast cancer samples were extracted with RNeasy Mini Kit (Qiagen), followed by quality control with RNA 6000 chips (Agilent Bioanalyzer). RNA species with RIN values $>$ 7 were sent to the Genomics Core of Yale Stem Cell Centre. Ribo-depleted RNA-Seq was conducted with 100 bp read length. The read count quantification and FPKM data were generated as above. The RNA-Seq reads of our samples will be deposited to GEO upon publishing of this manuscript. Tissue specificity

To analyze tissue specificity, Jensen-Shannon divergence score (JS score) was calculated from tumour and normal samples of each tissue, and the two distributions of JS scores were compared

following the method of Cabili et al [1]. Briefly, fragments per kilo bases of exons for per million mapped reads (FPKM) were first calculated from the normalized count data from each sample. Then the mean FPKM for each tissue type was calculated and log transformed. The vector e that represents the distribution of expression is given by:

$$e = \frac{log_2(FPKM + 1)}{\sum_{i=1}^{n} log_2(FPKM_i + 1)} \tag{9}$$

The JS_t score is the JS score for each tissue type t, calculated by the following:

$$JS_t(e,\ e^t) = \ 1 - \ \sqrt{H\left(e + e^t\right) - \ \frac{H\left(e\right) + H\left(e^t\right)}{2}} \tag{10}$$

Where H is the Shannon entropy and e^t is the hypothetical distribution when a lincRNA is expressed in only one tissue type:

$$e^t = (e^1, ..., e^i, ..., e^n),\ where\ e^i = \begin{cases} 1\ if\ \ i = t \\ 0\ if\ \ i \neq t \end{cases} \tag{11}$$

The JS score for a lincRNA is then defined as the maximum JS_t score across all tissue types.

4.3.2 Differential expression

Each of the 12 TCGA cancer datasets was tested for differential expression (DE) using DESeq [29, 30]. Statistically significant genes were selected with a FDR adjusted p-value threshold of 0.05 after Benjamini & Hochberg multiple hypothesis correction. As a result, six lincRNAs were discovered to be consistently upregulated or down-regulated in all twelve TCGA cancer datasets. These six lincRNAs were used subsequently for survival and pathway analysis.

4.3.3 Survival analysis

These six lincRNAs with pan-cancer diagnostic potential were examined for their association with patient survival among four types of TCGA cancer types. Note that these lincRNAs were initially selected as diagnostic biomarkers, but not prognostic biomarkers. The survival data from the four types TCGA cancers were obtained in two approaches. LUAD, LUSC and OV have relapse free survival information directly available from the TCGA data repository. The

fourth cancer type BRCA has overall survival data available, per the courtesy of Volinia et al [31]. Patients who did not have an event (death or tumour relapse, depending on the data set) during the study were considered as censored. The expression values of the six lincRNAs were used as predictors to fit a Cox-Proportional Hazards (Cox-PH) regression model, where the overall survival or disease free survival was the response variable. For each patient, a prognosis index (PI) score was generated from the Cox-PH model. The median PI score among all patients of the same cancer type was used as the threshold to dichotomize the patients into high vs. low risk groups, similar to other [32]. The log-rank p-value was then calculated to assess the statistically significant difference between the Kaplan-Meier curves of the high vs. low risk groups.

4.3.4 Tumour subtype classification and concordance between data types using NMF

Non-negative matrix factorization (NMF) method was used to classify tumour subtypes with lincRNA expression values. The optimal number of clusters was selected using the maximum cophenetic correlation. The lincRNA clustering results were then compared to those of other data types, using the method similar to Han et a [33]. The other data types from the TCGA include mRNA-Seq, mature microRNA-Seq, methylation and reverse phase protein array (RPPA) for each cancer typ [34], all obtained from the Broad institute Genomic Data Analysis Center (GDAC). The concordances from the chi-square tests between lincRNA and other data types were used to assess the correlations between clustering.

Additionally, lincRNA clustering was compared with another standard method, the PAM50 clustering [35], using the TCGA breast cancer samples. The correlation between these two clustering approaches was calculated using the concordance as mentioned above. Similarly, cluster correlation was computed for subtypes based on ER+/- information from the GSE58135 breast cancer dataset.

4.3.5 LincRNA sequence coding potential and homology characterization

To predict the coding potential of the sequences, iSeeRNA [34] and Coding-Potential Assessment Tool (CPAT) [36] were used. The two programs are trained on long non-coding RNAs to assess the coding potential of transcripts. For iSeeRNA, the coordinates of lincRNA transcripts and exons were used as inputs in the form of GFF files. For CPAT, lincRNA sequences were used as inputs in the form of fasta files. To test for homology between transcripts, NCBI's command

line BLAST+ suite [37] was used. Pairwise BLAST was performed on all isoforms of the six differentially expressed lincRNAs. We calculated the percentages of homology by the number of matching base pairs divided by the total number of base pairs in the query sequence. Due to the high homology between three of the discovered lincRNAs (PCAN-2, PCAN-3 and PCAN-5), downloaded RNA-Seq reads may have slight ambiguity in counting these lincRNA expression, since they were generated by TCGA using the Mapsplice alignment progra [26].

4.3.6 Quantitative RT-PCR (qRT-PCR) analysis

Total RNA from MDA-MB-231 and MCF-7 cell lines was isolated using RNeasy Mini Kit (Qiagen). Pooled total RNA from five healthy normal breast cancer patients was ordered from Biochain (Total RNA – Human Adult Normal Tissue 5 Donor Pool: Breast, catalog# R1234086-P). To match these healthy controls, total RNA was isolated from five in-house breast cancer patient samples.

High Capacity cDNA Reverse Transcription kit (Life Technologies, Thermo Scientific) was used for random-primed first-strand complementary DNA synthesis. Real time quantitative PCR (qPCR) was performed with SYBR Green (Life Technologies) with primers against selected linc RNAs (primer sequences are listed in Supplementary Table VI). Amplification and real time measurement of PCR products was performed with 7900HT Fast Real-Time PCR System (Life Technologies). The comparative Ct method [38] was used to quantify the expression levels of lincRNAs. Beta-glucuronidase (GUS) gene expression served as the internal control. GUS was selected as the internal control, as its expression level has been found to be comparable in range to the expression of linc RNAs and is stable in a wide variety of cancers [39, 40].

4.3.7 RNA interference

The siRNA oligos were synthesized by GE Dharmacon. The target sequences are as follows:

control siRNA #1: 5'-UGGUUUACAUGUCGACUAA-3'

control siRNA #2: 5'-UGGUUUACAUGUUGUGUGA-3'

control siRNA #3: 5'-UGGUUUACAUGUUUUCUGA-3'

control siRNA #4: 5'-UGGUUUACAUGUUUUCCUA-3'

lincRNA siRNA #1: 5'-UUCCUUUAGACCCAUUCUCUU-3'

lincRNA siRNA #2: 5'-PGAACCCACCACUGCUUCUC-3'

This lincRNA siRNA targets PCAN-2 and PCAN-3 lincRNAs. Cells were transfected in a 6-well plate format with siRNA oligos at 40nM (for cell proliferation assays) or 60nM (for migration assays) concentration, using DharmaFECT 1 Transfection Reagent (Dharmacon). The knockdown efficiency was determined by qRT-PCR 24 hours post transfection.

4.3.8 Cell growth and migration assays

Cell proliferation analysis was done using CellTiter-Glo Luminescent Cell Viability Assay Kit (Promega). Briefly, MDA-MB-231 cells were transfected in biological triplicates with siRNA constructs (control siRNA and linc RNA siRNA). After 24 hours, 400 cells of each condition were seeded in triplicates into 96-well plates and allowed to grow for another 48 hours. Cells number estimation at different time points was based on the quantification of the present ATP using SpectraMax Gemini XPS microplate reader (Molecular Devices). Cell migration was analysed using well established wound-healing assa [41]. Scratches in cell monolayer were made 30 hours post siRNA transfection (3 scratches in each of the 3 biological replicates). Cell migration was analysed by time-lapse microscopy using IX81 Olympus microscope, with 10x objective (for MDA-MB-231 cells) and 4x objective with additional 1.6x magnification (for MCF-7 cells). Images were taken every 5 minutes over time period of 24 hours. Migration rates and cell tracking were analysed using the Metamorph software.

4.4 Results

4.4.1 Overview of the workflow

To detect genes differentially expressed between healthy and tumour tissues, we employed a two-factor (cancer/normal, and source of samples) experimental design in which patients with tumour samples and matched normal sample were selected. This approach allowed sufficient statistical power by reducing the variation of data [42]. In total, we downloaded 1240 paired cancer and adjacent normal RNA-Seq samples in 12 different cancer types.

The 12 different cancer types include breast invasive carcinoma (BRCA), colon adenocarcinoma (COAD), head and neck squamous cell carcinoma (HNSC), kidney chromophobe (KICH), kidney renal clear cell carcinoma (KIRC), kidney renal papillary cell carcinoma (KIRP), liver hepatocellular carcinoma (LIHC), lung adenocarcinoma (LUAD), lung squamous cell carcinoma

(LUSC), prostate adenocarcinoma (PRAD), stomach adenocarcinoma (STAD) and thyroid carcinoma (THCA). Details on the number of samples in each cancer type, sequencing strategies, total mappable reads, and detected lincRNAs are listed in Supplementary Table I. For lincRNA genomic coordinates, we used the UCSC genome browser's "lincRNA transcript track", which is based on both the Broad Institute Human Body Map including the annotations of transcripts of uncertain coding potential (TUCP) [1]. We quantified lincRNA expression with normalized fragments per kilobase per million (FPKM) values. Computationally, we have performed various analyses to study the biological and clinical relevance of lincRNAs to pan-cancer, including differential expression (DE), tissue specificity and molecular subtype analyses, as well as construction and verification of the diagnostic and survival models (Supplementary figure 1). Experimentally, we have verified the gene expression differences of a panel of 6 lincRNAs, which have pan-cancer diagnostic biomarker potential. Most importantly, we demonstrated the phenotypic changes of two of the over-expressed lincRNAs by siRNA knockdown experiments in two breast cancer cell lines MCF-7 and MDA-MB-231.

4.4.2 The high tissue specificities of lincRNAs are diminished in cancers

To investigate the expression patterns of the lincRNA transcripts among different tissue types, we conducted principal component analysis (PCA) for lincRNA expression on adjacent normal and cancer samples separately from 12 TCGA datasets (Figure 1). As expected, the normal samples are clearly clustered by tissue type based on lincRNA expression (Figure 1a). However, the cancer samples become less separable by tissue type (Figure 1b). The less precise distinction of cancer samples in the PCA plot suggests a degree of de-differentiation of tumor cells. The possibility of confounding due to heterogeneity of tumours of the same type can be excluded, since the latter would lead to more spreading, rather than less spreading observed on the PCA plot. We therefore reason it as the loss of tissue specificity in cancers. Supporting this observation, the first three principal components of PCA account for less variance in cancer samples compared to those in the adjacent normal tissues, suggesting deregulation of lincRNAs in cancers (Figure 1). We replicated the same analysis for protein-coding genes between tumour and adjacent normal tissues, and found the same trend of losing tissue specificity in the tumour samples (Supplementary figure 2).

To further analyze the tissue specificity of lincRNAs, we calculated the tissue specificity scores (JS scores) as defined in Cabili et al[1], where a higher JS score indicates more tissue specificity. We compared the distributions of these JS scores in tumour and adjacent normal tissue, for both lincRNAs and RefSeq protein coding genes (Figure 2). Consistent with the PCA plots,

lincRNAs in cancer tissues are significantly less tissue specific than those in adjacent normal tissues (t-test, $p < 2.2e\text{-}16$) (Figure 2a, c and d). Moreover, in comparison with RefSeq protein coding genes (Figure 2b, e and f), lincRNAs have a much higher average JS score (t-test, $p < 2.2e\text{-}16$). Subsequently, we defined a subset of lincRNAs that are highly tissue specific with JS score greater than 0.75 and are expressed in at least 5% of the total normal samples (Supplementary Table II). To confirm that the tissue-specific lincRNAs defined by TCGA pan-cancer analysis are accurate, we then compared the tissue type assigned to lincRNAs by Cabili et al[1]to the tissue types assigned to the same lincRNAs based on the TCGA data. We observed statistically significant correlations (^{2}-test, all $p < 0.0001$) between the two studies in all tissue categories (Supplementary figure 3). In addition, we plot the tissue specific JS score for each tissue type (JS_t score) and plotted their distributions (Supplementary figure 4). As expected, significant amounts of lincRNA have zero JS scores, as many lincRNAs are not expressed in certain tissues.

4.4.3 LincRNA clustering accurately predicts molecular subtypes of tumours

Given the tissue specificity of lincRNAs, we hypothesized that lincRNAs can accurately separate tumours by molecular subtype. To identify a representative cancer type, we first used consensus non-negative matrix factorization (CNMF) to cluster the patient samples from each of the 12 types of cancer. We then calculated the correlations between the clustering result based on lincRNAs and those based on four other high-throughput data types: mRNA expression, micro-RNA expression, DNA methylation and reverse phase protein array (RPPA) obtained from the Broad Institute Genomic Data Analysis Center (GDAC) [24] (Broad, 2014) (Broad, 2014) The majority of lincRNA and GDAC clustering results are statistically significantly correlated (Figure 3a). As expected, lincRNA and mRNA expression are the most highly correlated among all four high-throughput data types. Among the 12 cancer datasets, the BRCA dataset has the best agreements between lincRNAs and the other data types. We therefore focused on the correlation between lincRNA and molecular subtypes in breast invasive carcinoma.

We first applied CNMF to the TCGA BRCA dataset and used cophenetic correlation [34] to determine the optimal cluster number to be 5, the same number of clusters as in PAM50 based classification. We then compared the result of CNMF clustering to PAM50 based subtypes, which include basal-like, HER2-enriched, luminal A, luminal B and normal-like subtypes [35] (Figure 3c). The concordance score based on the ^{2}-test is highly significant ($p < 2.2e\text{-}16$), and the overall accuracy to clinical types is 71.6%, as measured by rand measure, a metric for the percentage of agreement on a pair of samples belonging to the same group. Interestingly, the

first CNMF cluster has the strongest correlation with the basal-like subtype among all molecular subtypes, with an accuracy of 95% based on rand measure. Additionally, we examined the GSE58135 breast cancer dataset that has primary tumour samples in ER+/HER2- and triple negative subtypes (Figure 3b). The unsupervised CNMF clustering on these cancer samples yields highly accurate separation between ER+/HER2- and triple negative samples (^{2}-test p < 2.2e-16, and rand measure 84.5%). These results show that lincRNAs are well correlated with the molecular subtypes of tumours.

4.4.4 Transcriptome analysis reveals a pan-cancer panel of six lincRNAs

To seek a panel of lincRNAs as pan-cancer diagnostic biomarkers, we performed differential expression analysis on the above 12 TCGA datasets and detected thousands of differentially expressed lincRNAs in each TCGA dataset (Supplementary figure 5). Among them, six lincRNAs are consistently and significantly altered in all 12 cancers, with five of them being up-regulated and one down-regulated (Figure 4a, Supplementary figure 6 and Supplementary Table III). On the contrary, when we applied the same selection criteria to protein coding genes, we identified 47 mRNAs. The much larger number of mRNAs is presumably due to the less tissue specificity of mRNAs and more annotated mRNAs compared to lincRNAs at the time of investigation.

4.4.5 Analysis of known lincRNA markers

Several other lincRNAs, such as PCAT1, MALAT1, HOTAIR, have previously been reported to associate with a variety of cancers [13, 21, 22, 43]. We re-analyzed their expression in our pan-cancer data set (Supplementary figure 7). These three lincRNAs are not pan-cancer lincRNAs, but the TCGA results confirmed the previous findings based on several cancer types. PCAT1 was discovered in prostate cancer [22], and is indeed extremely significant in the TCGA PRAD data. MALAT1 is known to be primarily associated with liver cancer, lung cancer and kidney cancer [13], and it is recapitulated in the TCGA data. HOTAIR is also known to be highly upregulated in many different TCGA cancer types [44].

To confirm that the six lincRNAs are indeed associated with pan-cancers, we processed additional 833 samples from a wide range of resources including three public RNA-Seq datasets and eleven microarray datasets (Supplementary Table I). All three public RNA-Seq datasets (GSE58135 breast cancer, GSE50760 colon cancer, and GSE25599 liver cancer) show consistent directions of fold change for all six lincRNAs (Figure 4b). Although the microarray platforms are not designed to detect lincRNAs, some probes are nevertheless overlapped with non-coding

RNAs as shown by others [45], and thus they can be another source of empirical verification. Among the various microarray platforms examined, 24 of the 29 microarray probe sets have the same overall directions of fold changes as those in the RNA-Seq datasets (Supplementary figure 8). Moreover, the expression levels of the six lincRNAs in 28 breast cancer cell lines from the GSE58135 dataset and 5 breast cancer cell lines from the Cancer Cell Line Encyclopedia (CCLE) are all comparable with those from the TCGA BRCA samples (Supplementary figure 9), further supporting the robustness of these lincRNAs as potential pan-cancer biomarkers.

To verify this lincRNA panel experimentally, we performed additional RNA-Seq and qPCR experiments on our own breast cancer samples. First, we sequenced fresh frozen primary tumour samples from 10 individual patients using the ribosomal depletion RNA-Seq method. We then compared them to normal breast tissue RNA-Seq data from GEO (GSE52194, GSE45326 and GSE30611). All six lincRNAs have the same trends of changes as in the other GEO RNA-Seq datasets (Figure 4c) and five of them are significantly differentially expressed. We followed up with the qPCR validation and designed seven PCR primer pairs for selected transcripts in the lincRNA panel (supplementary table IV). The qPCR results in pooled breast tumour samples (n=5), pooled normal breast samples (n=5) and MCF-7 cell lines are shown in Figure 4d. In all cases, the expression levels show statistically significant differential expression in the same directions as the RNA-Seq data, both between primary tumour and normal sample pools and between normal and MCF-7 cancer cell lines.

4.4.6 Sequence features among the six-lincRNA biomarkers

To confirm the non-coding nature of the lincRNA transcripts, we used the iSeeRN [34] and Coding-Potential Assessment Tool (CPAT) [36]. Both programs are specifically trained on long non-coding RNAs to assess the non-coding potential of RNA transcripts. Out of the 52 isoforms from the lincRNA panel, iSeeRNA predicted 49 to be non-coding. For the three transcripts that are ambiguous, we used a second tool, CPAT, to obtain further evidence for the coding or non-coding nature of these transcripts. CPAT classifies all three of them as non-coding RNAs. In contrast, both CPAT and iSeeRNA correctly classified all isoforms of house-keeping genes GUS and GAPDH as protein coding. Overall, both programs provide strong evidence for the non-coding nature of the six lincRNAs (Supplementary Table V).

To examine the relationship between the six lincRNAs, we first checked the correlations of their expression values in all TCGA samples. Three of the lincRNAs, PCAN-2, PCAN-3 and PCAN-5, are highly correlated with spearman correlation coefficients of approximately 0.92 between them (Supplementary figure 10). The high correlations among expression prompted us to check

if sequence similarities exist. Thus, we tested the pairwise homology among all transcripts of the six lincRNAs, using NCBI's BLAST+ suit [37] (Supplementary figure 11). Indeed, the three lincRNAs mentioned above are highly homologous, and some of the annotated transcripts are 99% identical. Two of the lincRNAs, PCAN-2 and PCAN-3, are in the tandem locations on chromosome 14 and the third lincRNA PCAN-5 is located on chromosome 22, suggesting potential gene duplication events from a common origin.

4.4.7 The lincRNA biomarker panel robustly and accurately predicts pan cancers

To quantitatively assess the value of the six lincRNAs as pan-cancer diagnostic biomarkers, we built a classification model upon them (Figure 5a). First, we split the TCGA pan-cancer data into 80% training and 20% holdout testing sets. Given that some lincRNAs are highly correlated (Supplementary figure 10) and thus potentially redundant as biomarker predictors, we used correlation feature selection (CFS) method to select the most relevant and least redundant subset of lincRNAs among them. As a result, five of the lincRNAs were chosen: PCAN-1, PCAN-2, PCAN-3, PCAN-4, and PCAN-6.

We then compared the classification results on the training dataset using four widely used machine-learning algorithms: Random Forest (RF), Linear Support Vector Machines (LSVM), Gaussian Support Vector Machines (GSVM) and Logistic Regression with L2 regularization (L2-LR). As shown by the receiver operator characteristics (ROC) curves on the TCGA training data set, RF has the best AUC of 0.947 (95% confidence interval, or CI: 0.9343-0.9603) on the training data among the four methods (Supplementary figure 12). We thus selected the RF model to test the classification performance on additional 496 samples from the hold out test set. As expected, the trained RF model has very similar prediction result on the TCGA hold-out testing set, with an AUC=0.947, sensitivity=0.817 and specificity=0.970 (Figure 5d).

To further verify the robustness of the five-lincRNA panel, we tested the TCGA data based RF model on four independent RNA-Seq datasets: GSE58135 breast cancer, GSE50760 colon cancer, GSE25599 liver cancer and our breast cancer dataset (Figure 5b, c and d). Impressively, this model predicts the other four independent data sets very well, with AUCs of 0.972 (95% CI: 0.95-0.9946), 0.841(95% CI: 0.6875-0.9946), 0.970 (95% CI: 0.9108-1) and 0.950 (95% CI: 0.867-1) for GSE58135, GSE50760, GSE25599 and our dataset, respectively (Figure 5c and d). Other model evaluation metrics including Sensitivity, Specificity, Precision, Matthew's Correlation Coefficient, F-score and Accuracy in the validation datasets further demonstrate the excellent

performance of the model (Supplementary Table VI). We therefore conclude that the panel of six lincRNAs are potential biomarkers for pan-cancer diagnosis.

4.4.8 The lincRNA panel is associated with prognosis in cancer patients

Although the six lincRNAs were detected as potential diagnosis markers for pan-cancer, we were curious if they might be associated with the prognosis of cancer patients as well. Thus we performed survival analysis on 1201 samples from four TCGA datasets: namely BRCA, LUAD, LUSC datasets, and additionally the TCGA ovarian cancer (OV) dataset which was not used in the lincRNA signature discovery phase due to lack of normal samples (Supplementary figure 13). Since only overall survival information is available in TCGA in BRCA and OV datasets, we fit the overall survival with Cox-PH regression models and categorized the patient risks by prognosis index (PI) [32]. The resulting Kaplan-Meier survival curves show that the lincRNA panel is able to separate patients into higher and lower risk groups by median PI, with log-rank tests p-values of 0.012 and 0.010 for BRCA and grade 3 OV, respectively (Supplementary figure 13a and b). On the other hand, the more preferable relapse free survival (RFS) in LUAD and LUSC datasets are available, thus we fit RFS with Cox-PH models, and obtained significant p-values of 0.0416 and 0.013 for differential survivals of LUAD and LUSC samples, respectively (Supplementary figure 13c and d). In summary, although the lincRNA panel was not purposely discovered as prognosis markers but rather diagnostic markers, their expression values are associated with the prognosis outcomes in various types of cancers.

4.4.9 Biological relevance of lincRNAs explored by cell culture experiments

To explore the relationship between the lincRNAs panel and tumourigenic phenotypes, we conducted experiments using two breast cancer and colon cancer cell lines as examples. Given the extremely high homology between PCAN-2 and PCAN-3, we intentionally designed siRNAs that target both of them so as to observe phenotypes. In non-aggressive MCF-7 and highly metastatic MDA-MB-231 cell lines, we efficiently knocked down two lincRNAs PCAN-2 and PCAN-3 (Figure 6a). Transient knockdown allowed us to analyse cell proliferation and cell migration rate. Interestingly, the growth rate of fast proliferating MDA-MB-231 cells significantly decreased upon transfection with lincRNAs siRNA (Figure 6b). To assess cell migration rates we employed the well-established wound-healing assay and followed the cell movement with time-lapse microscoopy over the time of 24 hours. As expected, the migration rate was significantly inhibited upon lincRNAs knockdown (Figure 6c, d). The effect of lincRNA down-regulation on cell migration was more pronounced in a highly aggressive MDA-MB-231 cell line

(0.349 versus 0.059 mm over 24 hours for control and lncRNA siRNA, respectively) but it was also observed in much slower migrating MCF-7 cells (0.127 versus 0.096 mm over 24 hours for control and lncRNA siRNA, respectively). We repeated the cell migration experiment on MDA-MB-231 with another less effective siRNA, and observed similar significantly slower ($P<0.0001$) migrating rate (Supplementary figure 14).

Furthermore, we repeated these experiments in another HCT116 colon cancer cell line with the more efficient siRNA (supplementary figure 15). Using the same experimental procedures, we observed significant differences in both cell proliferation ($p<0.0001$) and migration ($p=0.036$), between the lincRNA knockdown and the siRNA scrambled control. These results suggest that down-regulation of cancer cell abundant PCAN-2 and PCAN-3 lincRNAs weakens the typical cancer phenotypic features, such as proliferation and migration.

4.5 Discussion

Since 2012, a community effort has launched towards TCGA pan-cancer analysis across many different tumour type [33, 46], where the main focus has been the mutational landscap [47]. Pan-Cancer Initiative aims to enable the discovery of novel intervention strategies that can be tested clinically, including developing novel adaptive biomarker-based clinical trials that cross boundaries between tumour type [24]. One can expect that in the future, a pan-cancer screening biomarker panel from blood or other body fluids could become a useful, routine, and economical screening too [24] applied before the patients have typical cancer symptoms that indicate late-stage character of the disease. Once an individual is identified as high-risk in the test, he or she can be followed up with more confirmative tests, such as imaging scanning. In the field of cancer biomarkers, although many lincRNAs and other lncRNAs have recently been implicated in cancer initiation and progressio [33, 48, 49], the clinical potential of lincRNAs remains under-explored across different tumour types. In this study, our goals were to (1) depict the landscape of lincRNAs in pan-cancers, (2) demonstrate their relevance to clinical outcomes, such as tumour subtype, diagnosis and patient survival; and (3) explore the utilities of lincRNAs as pan-cancer diagnostic biomarkers.

Towards these goals, we have performed a new dimension of pan-cancer analysis using the lincRNA transcriptome. In total, we analyzed 3354 patient RNA-Seq samples from 12 types of cancers in TCGA (13 including OV in survival analysis) as well as an additional 15 independent datasets (three RNA-Seq datasets from GEO, one in-house RNA-Seq breast cancer dataset and 11 microarray datasets from GEO). To our knowledge, this study is the most comprehensive

endeavour to analyze lincRNAs in the context of pan-cancer. By systematically analyzing 12 types of RNA-Seq datasets in TCGA, we show that lincRNAs are more tissue specific than protein-coding genes. The loss of tissue specificity due to cancer is greater for lincRNAs compared to protein-coding genes. This suggests that lincRNAs can potentially be more sensitive biomarkers than protein coding genes. In addition, unsupervised clustering results of lincRNAs demonstrate significant correlations with molecular subtypes. CNMF clustering based on lincRNAs almost perfectly divided the Triple Negative and ER+/Her2- breast cancers into distinct groups in GSE58135 data set. Furthermore, CNMF clustering of TCGA BRCA samples detected 5 distinct clusters that highly correspond to the five widely used molecular subtypes based on the PAM50 signatures.

Although others have suggested that lincRNAs have potential as biomarker [22, 34], our study is the first to pinpoint a promising six-lincRNA pan-cancer diagnostics panel quantitatively, rigorously and robustly. Despite all the potential issues including population heterogeneity and sample size limitation in high throughput dataset [50], the six-lincRNA biomarker model performs well overall with AUCs ranging from 0.972 to 0.841. Moreover, we verify the alteration of these lincRNAs with eleven additional microarray gene expression data sets. Our most unexpected finding is that the six lincRNA diagnostic signature is also associated with the survival prognosis of cancer patients, based on the TCGA datasets (BRCA, OV, LUAD and LUSC). Furthermore, we have demonstrated that the lincRNAs have biological functions, by knocking-down experiments on two of them, PCAN-2 and PCAN-3. Our preliminary results indicate that downregulation of only two out of six panel lincRNAs is sufficient to partially revert some of the typical physiological hallmarks of cancer cells including fast proliferation and more importantly, migration.

Developing a pan-cancer biomarker model based on the lincRNA signatures could be very significant clinically, providing complementary values to protein-coding gene based biomarker panels. We plan to continue our translational investigations in this direction. Yet our next challenge is to understand how each of the identified lincRNA biomarkers function in tumourigenesis and progression. Although lincRNAs do not encode proteins, it's clear that they play important roles in cellular biology. Currently, multiple hypotheses exist on how lincRNAs regulate cellular functions [2], which include functioning as scaffold structure [23, 51], sponge of small regulatory RNA [10, 11] or direct interaction with proteins to modulate localization and activit [52]. To better understand the phenotypic effects of the six lincRNAs, we will proceed with experiments that address the physiological functions of these lincRNAs as well as molecular mechanisms by which they promote tumourigenesis and/or malignancy. We are aware that the repertoire of lincRNAs is evolving and thus we may miss some newly identified lincRNAs, such as reported

recentl [48]. However, given the fact that the six lincRNAs in this report have reached very high and robust accuracy in pan-cancer data, the addition of other new lincRNAs is expected to not increase the robustness of the current panel.

In summary, our initial pan-cancer analysis has demonstrated that lincRNAs accurately classify cancer subtypes through supervised as well as unsupervised methods. The panel of six lincRNAs is a highly accurate diagnostic biomarker signature with additional prognostic value. These results highlight lincRNAs as a new paradigm for actionable pan-cancer diagnosis and prognosis.

4.6 Acknowledgements

This research was supported by grants K01ES025434 awarded by NIEHS through funds provided by the trans-NIH Big Data to Knowledge (BD2K) initiative (www.bd2k.nih.gov), P20 COBRE GM103457 awarded by NIH/NIGMS, and Hawaii Community Foundation Medical Research Grant 14ADVC-64566 to L.X. Garmire. B. Fogelgren is supported by awards from NIH (1K01DK087852, R03DK100738, and P20GM103456-06A1-8293); the March of Dimes (#5-FY14-56); Hawaii Community Foundation (12ADVC-51347); University of Alabama at Birmingham HepatoRenal Fibrocystic Disease Core Center (5P30DK074038), and RCMI-BRIDGES at the University of Hawaii (5G12MD007601). The UHCC GSR is supported by the NCI P-30 grant CA071789-15. We would like to thank Dr Joe Ramos and Paul Anastasiadis for providing breast cancer cell lines and the UHCC Microscopy and Imaging Shared Resource for using the facility.

4.7 Author contributions

LXG and TC envisioned the project and designed the work. TC conducted the data analysis, with assistance from SH and XZ. TC and LXG wrote the manuscript. SY and HY provided the UHCC RNA samples, and RF helped to sequence the UHCC breast cancer samples. MT, TC and LXG designed the qPCR primers and KP, JM, MT and BF collaborated on cell culture and qPCR validations. All authors have read, revised and approved the final manuscript.

4.8 Competing financial interests

The author(s) declare no competing financial interests.

References

1. Cabili, M. N., Trapnell, C., Goff, L., Koziol, M., Tazon-Vega, B., Regev, A. & Rinn, J. L. Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. *Genes & Development* **25,** 1915–1927. ISSN: 0890-9369, 1549-5477 (2011).
2. Ching, T., Masaki, J., Weirather, J. & Garmire, L. X. Non-coding yet non-trivial: a review on the computational genomics of lincRNAs. *BioData mining* **8,** 1. ISSN: 1756-0381 (2015).
3. Garmire, L. X., Garmire, D. G., Huang, W., Yao, J., Glass, C. K. & Subramaniam, S. A global clustering algorithm to identify long intergenic non-coding RNA-with applications in mouse macrophages. *PloS one* **6,** e24051. ISSN: 1932-6203 (2011).
4. Sun, K., Chen, X., Jiang, P., Song, X., Wang, H. & Sun, H. iSeeRNA: identification of long intergenic non-coding RNA transcripts from transcriptome sequencing data. *BMC genomics* **14,** S7. ISSN: 1471-2164 (2013).
5. Consortium, E. P. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489,** 57–74. ISSN: 0028-0836 (2012).
6. McHugh, C. A., Russell, P. & Guttman, M. Methods for comprehensive experimental identification of RNA-protein interactions. *Genome Biol* **15,** 203 (2014).
7. Ma, H., Hao, Y., Dong, X., Gong, Q., Chen, J., Zhang, J. & Tian, W. Molecular mechanisms and function prediction of long noncoding RNA. *The Scientific World Journal* **2012** (2012).
8. Kandoth, C. *et al.* Mutational landscape and significance across 12 major cancer types. *Nature* **502,** 333–9. ISSN: 1476-4687 (Electronic) 0028-0836 (Linking) (2013).
9. Liang, C. C., Park, A. Y. & Guan, J. L. In vitro scratch assay: a convenient and inexpensive method for analysis of cell migration in vitro. *Nat Protoc* **2,** 329–33. ISSN: 1750-2799 (Electronic) 1750-2799 (Linking) (2007).
10. Ling, H., Spizzo, R., Atlasi, Y., Nicoloso, M., Shimizu, M., Redis, R. S., Nishida, N., GafA, R., Song, J. & Guo, Z. CCAT2, a novel noncoding RNA mapping to 8q24, underlies metastatic progression and chromosomal instability in colon cancer. *Genome research* **23,** 1446–1461. ISSN: 1088-9051 (2013).
11. Rubie, C., Kempf, K., Hans, J., Su, T., Tilton, B., Georg, T., Brittner, B., Ludwig, B. & Schilling, M. Housekeeping gene variability in normal and cancerous colorectal, pancreatic, esophageal, gastric and hepatic tissues. *Molecular and cellular probes* **19,** 101–109. ISSN: 0890-8508 (2005).

12. Wilks, C., Cline, M. S., Weiler, E., Diehkans, M., Craft, B., Martin, C., Murphy, D., Pierce, H., Black, J. & Nelson, D. The Cancer Genomics Hub (CGHub): overcoming cancer through the power of torrential data. *Database* **2014,** bau093. ISSN: 1758-0463 (2014).

13. Iyer, M. K., Niknafs, Y. S., Malik, R., Singhal, U., Sahu, A., Hosono, Y., Barrette, T. R., Prensner, J. R., Evans, J. R. & Zhao, S. The landscape of long noncoding RNAs in the human transcriptome. *Nature Genetics.* ISSN: 1061-4036 (2015).

14. Trapnell, C., Williams, B. A., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M. J., Salzberg, S. L., Wold, B. J. & Pachter, L. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature biotechnology* **28,** 511–515 (2010).

15. Tripathi, V., Ellis, J. D., Shen, Z., Song, D. Y., Pan, Q., Watt, A. T., Freier, S. M., Bennett, C. F., Sharma, A. & Bubulya, P. A. The nuclear-retained noncoding RNA MALAT1 regulates alternative splicing by modulating SR splicing factor phosphorylation. *Molecular cell* **39,** 925–938. ISSN: 1097-2765 (2010).

16. Gupta, R. A., Shah, N., Wang, K. C., Kim, J., Horlings, H. M., Wong, D. J., Tsai, M.-C., Hung, T., Argani, P. & Rinn, J. L. Long non-coding RNA HOTAIR reprograms chromatin state to promote cancer metastasis. *Nature* **464,** 1071–1076. ISSN: 0028-0836 (2010).

17. Prensner, J. R., Iyer, M. K., Balbin, O. A., Dhanasekaran, S. M., Cao, Q., Brenner, J. C., Laxman, B., Asangani, I. A., Grasso, C. S. & Kominsky, H. D. Transcriptome sequencing across a prostate cancer cohort identifies PCAT-1, an unannotated lincRNA implicated in disease progression. *Nature biotechnology* **29,** 742–749. ISSN: 1087-0156 (2011).

18. Brockdorff, N., Ashworth, A., Kay, G. F., Cooper, P., Smith, S., McCabe, V. M., Norris, D. P., Penny, G. D., Patel, D. & Rastan, S. Conservation of position and exclusive expression of mouse Xist from the inactive X chromosome. *Nature* **351,** 329–331. ISSN: 0028-0836 (1991).

19. Menor, M., Ching, T., Zhu, X., Garmire, D. & Garmire, L. X. mirMark: a site-level and UTR-level classifier for miRNA target prediction. *Genome biology* **15,** 500. ISSN: 1465-6906 (2014).

20. Wang, L., Park, H. J., Dasari, S., Wang, S., Kocher, J.-P. & Li, W. CPAT: Coding-Potential Assessment Tool using an alignment-free logistic regression model. *Nucleic acids research* **41,** e74–e74. ISSN: 0305-1048 (2013).

21. Ge, X., Chen, Y., Liao, X., Liu, D., Li, F., Ruan, H. & Jia, W. Overexpression of long noncoding RNA PCAT-1 is a novel biomarker of poor prognosis in patients with colorectal cancer. *Medical oncology* **30,** 1–6. ISSN: 1357-0560 (2013).

22. Penny, G. D., Kay, G. F., Sheardown, S. A., Rastan, S. & Brockdorff, N. Requirement for Xist in X chromosome inactivation. *Nature* **379,** 131–137. ISSN: 0028-0836 (1996).

23. Liao, Y., Smyth, G. K. & Shi, W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **30,** 923–930. ISSN: 1367-4803 (2014).

24. Cancer Genome Atlas Research, N., Weinstein, J. N., Collisson, E. A., Mills, G. B., Shaw, K. R., Ozenberger, B. A., Ellrott, K., Shmulevich, I., Sander, C. & Stuart, J. M. The Cancer Genome Atlas Pan-Cancer analysis project. *Nat Genet* **45,** 1113–20. ISSN: 1546-1718 (Electronic) 1061-4036 (Linking) (2013).

25. Weinstein, J. N., Collisson, E. A., Mills, G. B., Shaw, K. R. M., Ozenberger, B. A., Ellrott, K., Shmulevich, I., Sander, C., Stuart, J. M. & Network, C. G. A. R. The cancer genome atlas pan-cancer analysis project. *Nature genetics* **45,** 1113–1120. ISSN: 1061-4036 (2013).

26. Volinia, S. & Croce, C. M. Prognostic microRNA/mRNA signature from the integrated analysis of patients with invasive breast cancer. *Proc Natl Acad Sci U S A* **110,** 7413–7. ISSN: 1091-6490 (Electronic) 0027-8424 (Linking) (2013).

27. Liao, Q., Liu, C., Yuan, X., Kang, S., Miao, R., Xiao, H., Zhao, G., Luo, H., Bu, D. & Zhao, H. Large-scale prediction of long non-coding RNA functions in a coding–non-coding gene co-expression network. *Nucleic acids research* **39,** 3864–3878. ISSN: 0305-1048 (2011).

28. Liao, Y., Smyth, G. & Shi, W. featureCounts: an efficient general-purpose read summarization program. *arXiv* **1305,** 16 (2013).

29. Loewen, G., Zhuo, Y., Zhuang, Y., Jayawickramarajah, J. & Shan, B. lincRNA HOTAIR as a novel promoter of cancer progression. *Journal of Cancer Research Updates* **3,** 134–140. ISSN: 1929-2279 (2014).

30. Love, M., Anders, S. & Huber, W. Differential analysis of RNA-Seq data at the gene level using the DESeq2 package. *Bioconductor* (2013).

31. Vitiello, M., Tuccoli, A. & Poliseno, L. Long non-coding RNAs in cancer: implications for personalized therapy. *Cellular Oncology,* 1–12. ISSN: 2211-3428 (2014).

32. Harrow, J. *et al.* GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Res* **22,** 1760–74. ISSN: 1549-5469 (Electronic) 1088-9051 (Linking) (2012).

33. Han, L., Yuan, Y., Zheng, S., Yang, Y., Li, J., Edgerton, M. E., Diao, L., Xu, Y., Verhaak, R. G. & Liang, H. The Pan-Cancer analysis of pseudogene expression reveals biologically and clinically relevant tumour subtypes. *Nature communications* **5** (2014).

34. Salmena, L., Poliseno, L., Tay, Y., Kats, L. & Pandolfi, P. P. A ceRNA Hypothesis: The Rosetta Stone of a Hidden RNA Language? *Cell* **146,** 353–358. ISSN: 0092-8674 (2011).

35. Cancer Genome Atlas, N. Comprehensive molecular portraits of human breast tumours. *Nature* **490,** 61–70. ISSN: 1476-4687 (Electronic) 0028-0836 (Linking) (2012).

36. Wang, K., Singh, D., Zeng, Z., Coleman, S. J., Huang, Y., Savich, G. L., He, X., Mieczkowski, P., Grimm, S. A. & Perou, C. M. MapSplice: accurate mapping of RNA-seq reads for splice junction discovery. *Nucleic acids research,* gkq622. ISSN: 0305-1048 (2010).

37. Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K. & Madden, T. L. BLAST+: architecture and applications. *BMC bioinformatics* **10,** 421. ISSN: 1471-2105 (2009).

38. Liu, K., Yan, Z., Li, Y. & Sun, Z. Linc2GO: a human LincRNA function annotation resource based on ceRNA hypothesis. *Bioinformatics* **29,** 2221–2222. ISSN: 1367-4803 (2013).

39. Habel, L. A., Shak, S., Jacobs, M. K., Capra, A., Alexander, C., Pho, M., Baker, J., Walker, M., Watson, D. & Hackett, J. A population-based study of tumor gene expression and risk of breast cancer death among lymph node-negative patients. *Breast Cancer Res* **8,** R25 (2006).

40. Rinn, J. L., Kertesz, M., Wang, J. K., Squazzo, S. L., Xu, X., Brugmann, S. A., Goodnough, L. H., Helms, J. A., Farnham, P. J. & Segal, E. Functional Demarcation of Active and Silent Chromatin Domains in Human HOX Loci by Noncoding RNAs. *Cell* **129,** 1311–1323. ISSN: 0092-8674 (2007).

41. Kowalczyk, M. S., Higgs, D. R. & Gingeras, T. R. Molecular biology: RNA discrimination. *Nature* **482,** 310–311. ISSN: 0028-0836 (2012).

42. Ching, T., Huang, S. & Garmire, L. X. Power analysis and sample size estimation for RNA-Seq differential expression. *RNA* **20,** 1684–96. ISSN: 1469-9001 (Electronic) 1355-8382 (Linking) (2014).

43. Chiyomaru, T., Fukuhara, S., Saini, S., Majid, S., Deng, G., Shahryari, V., Chang, I., Tanaka, Y., Enokida, H. & Nakagawa, M. Long non-coding RNA HOTAIR is targeted and regulated by miR-141 in human cancer cells. *Journal of Biological Chemistry* **289,** 12550–12565. ISSN: 0021-9258 (2014).

44. Livak, K. J. & Schmittgen, T. D. Analysis of relative gene expression data using real-time quantitative PCR and the 2-delta-delta-CT method. *methods* **25,** 402–408. ISSN: 1046-2023 (2001).

45. Du, Z., Fei, T., Verhaak, R. G., Su, Z., Zhang, Y., Brown, M., Chen, Y. & Liu, X. S. Integrative genomic analyses reveal clinically relevant long noncoding RNAs in human cancer. *Nat Struct Mol Biol* **20,** 908–13. ISSN: 1545-9985 (Electronic) 1545-9985 (Linking) (2013).

46. Weakley, S. M., Wang, H., Yao, Q. & Chen, C. Expression and function of a large non-coding RNA gene XIST in human cancer. *World journal of surgery* **35,** 1751–1756. ISSN: 0364-2313 (2011).

47. Ji, P., Diederichs, S., Wang, W., Boeing, S., Metzger, R., Schneider, P. M., Tidow, N., Brandt, B., Buerger, H. & Bulk, E. MALAT-1, a novel noncoding RNA, and thymosin beta-4 predict metastasis and survival in early-stage non-small cell lung cancer. *Oncogene* **22,** 8031–8041. ISSN: 0950-9232 (2003).

48. Huang, S., Yee, C., Ching, T., Yu, H. & Garmire, L. X. A Novel Model to Combine Clinical and Pathway-Based Transcriptomic Information for the Prognosis Prediction of Breast Cancer. *PLoS computational biology* **10,** e1003851. ISSN: 1553-7358 (2014).

49. Ulitsky, I. & Bartel, D. P. lincRNAs: genomics, evolution, and mechanisms. *Cell* **154,** 26–46. ISSN: 0092-8674 (2013).

50. Berrar, D., Bradbury, I. & Dubitzky, W. Avoiding model selection bias in small-sample genomic datasets. *Bioinformatics* **22,** 1245–1250. ISSN: 1367-4803 (2006).

51. Khalil, A. M., Guttman, M., Huarte, M., Garber, M., Raj, A., Morales, D. R., Thomas, K., Presser, A., Bernstein, B. E. & van Oudenaarden, A. Many human large intergenic noncoding RNAs associate with chromatin-modifying complexes and affect gene expression. *Proceedings of the National Academy of Sciences* **106,** 11667–11672. ISSN: 0027-8424 (2009).

52. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-Seq data with DESeq2. *bioRxiv* (2014).

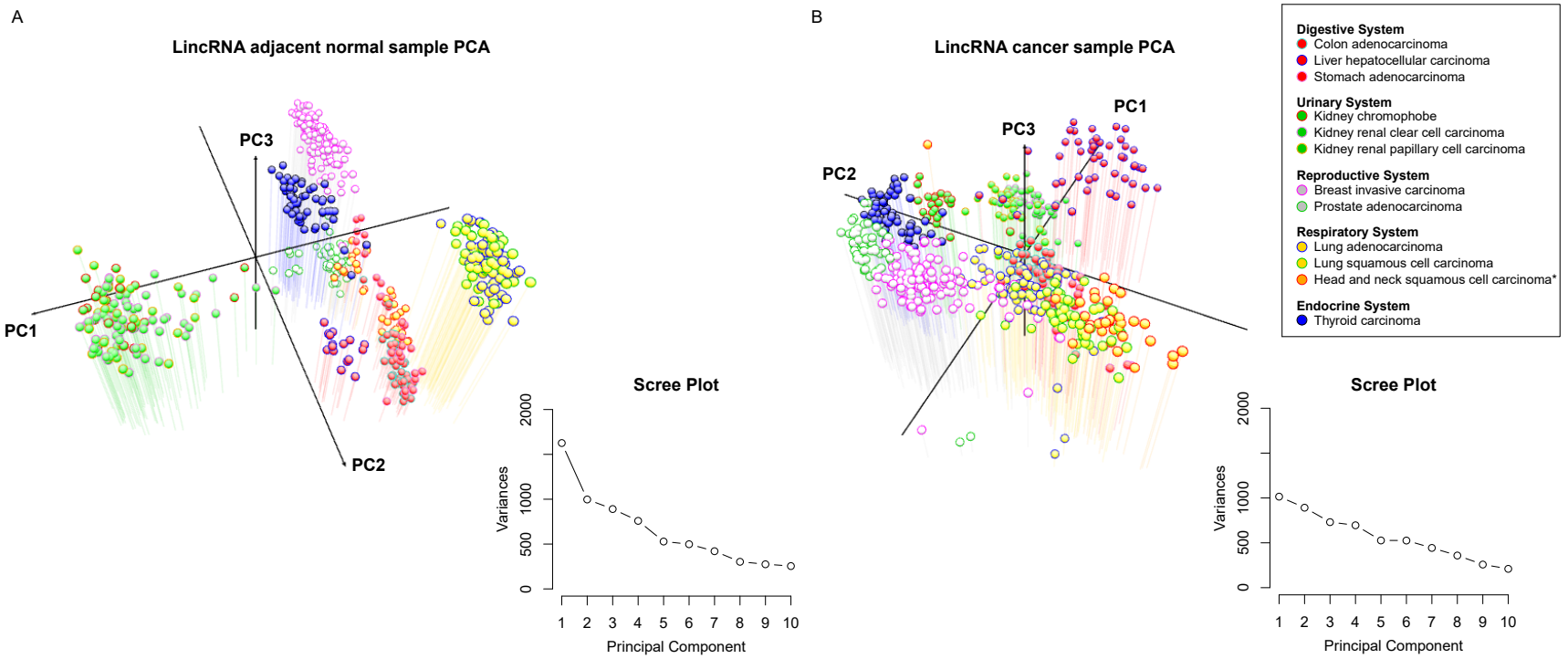


Figure 1: Principal component analysis of lincRNA expression in 12 TCGA datasets. The first three principal components (PCs) were plotted using the log FPKM values of lincRNA expression in (a) normal adjacent tissue and (b) cancer samples. The variances associated with each of the first 10 principal components are plotted alongside each graph (Scree Plot).

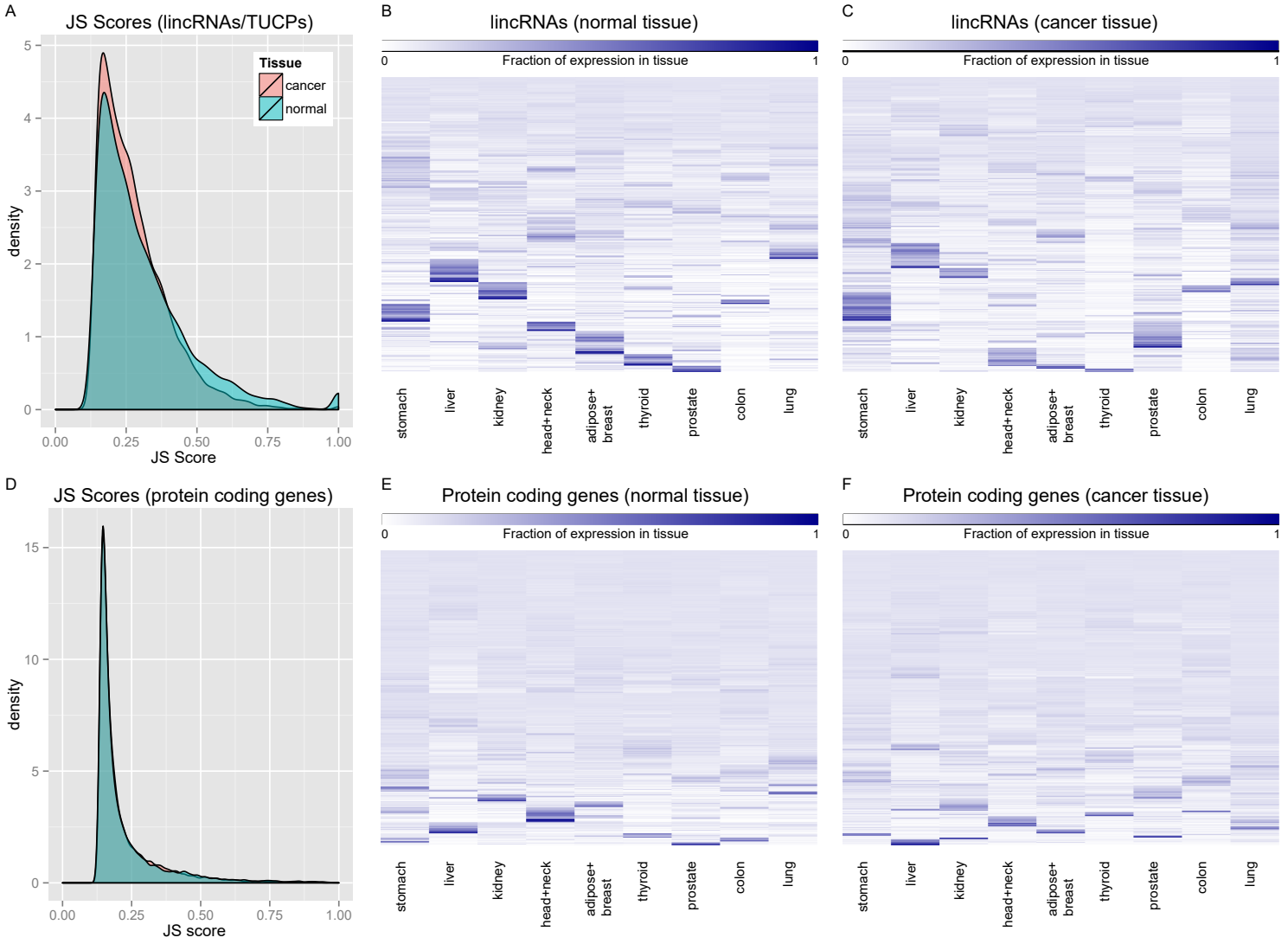


Figure 2: Tissue specificities of lincRNAs and protein coding genes. (a-b) Maximum JS scores for were used to measure tissue specificity in primary tumours and adjacent normal samples, based on either lincRNAs (a) or protein coding genes (b). A value of 1 indicates that the lincRNA is expressed in only one tissue. (c-f) Fractional expression of lincRNAs or protein coding genes in each tissue was plotted in the adjacent normal or cancer samples.

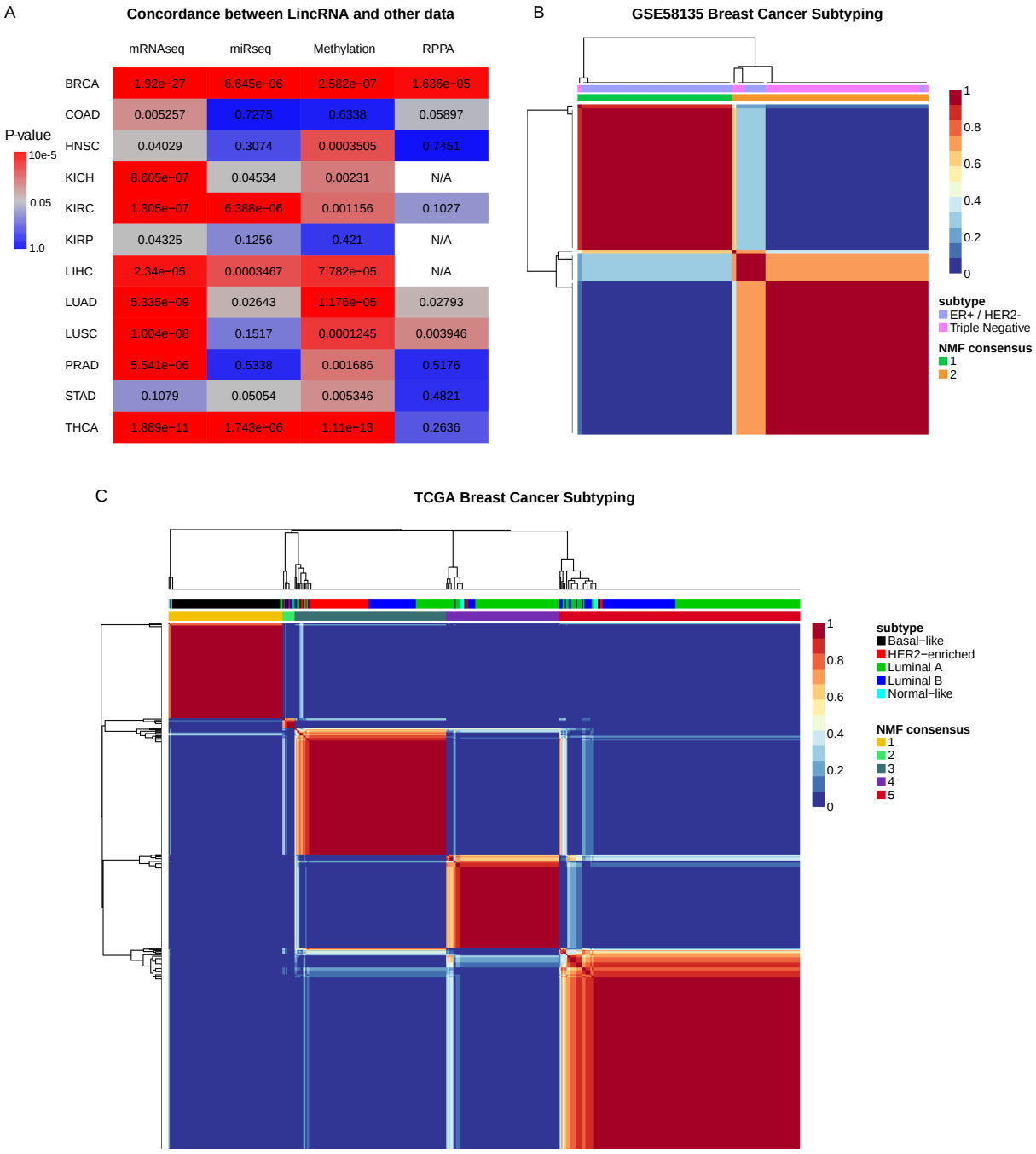


Figure 3: Correlation of lincRNAs with other data types and cancer subtypes. (a) The concordance between clustering results of lincRNAs and other high throughput data types in TCGA based on chi-square statistical test. (b) CNMF was used to determine the clustering of lincRNAs in the GSE58135 Breast Cancer dataset. The concordance of the clustering with the tumour subtypes in the dataset is significant (chi-square, $p < 2.2e\text{-}16$). (c) CNMF was used to determine the clustering of lincRNA in the TCGA BRCA dataset. The concordance of the CNMF clustering with the tumour subtypes in the dataset is significant (chi-square, $p < 2.2e\text{-}16$).

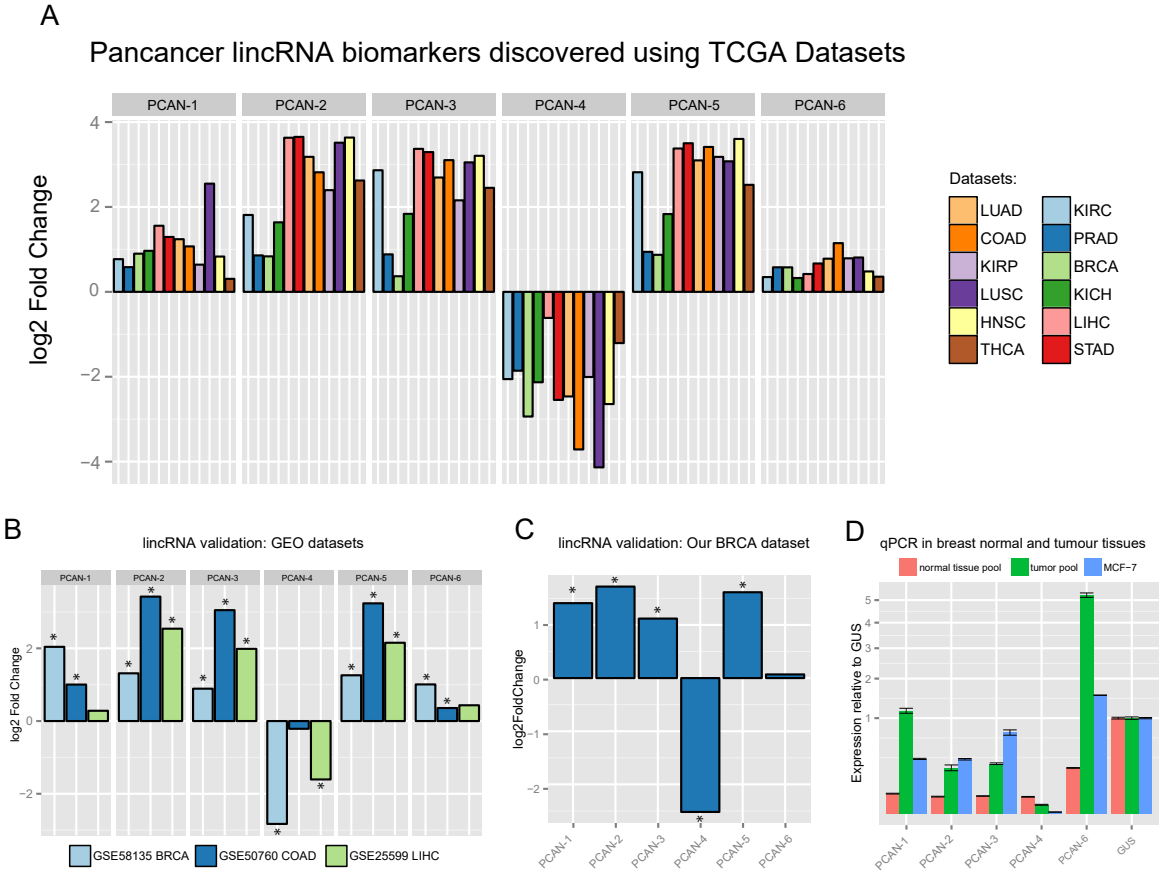


Figure 4: Differentially expressed pan-cancer lincRNAs. (a) Six lincRNAs are consistently differentially expressed in 12 TCGA datasets. Each of the six lincRNAs shown is either significantly upregulated or significantly downregulated across the various cancers. The six lincRNAs in three independent RNA-Seq datasets from GEO (b), our own breast cancer dataset (c) and qPCR of pooled 5 normal tissues, pooled 5 tumours and the MCF-7 cell line (d).

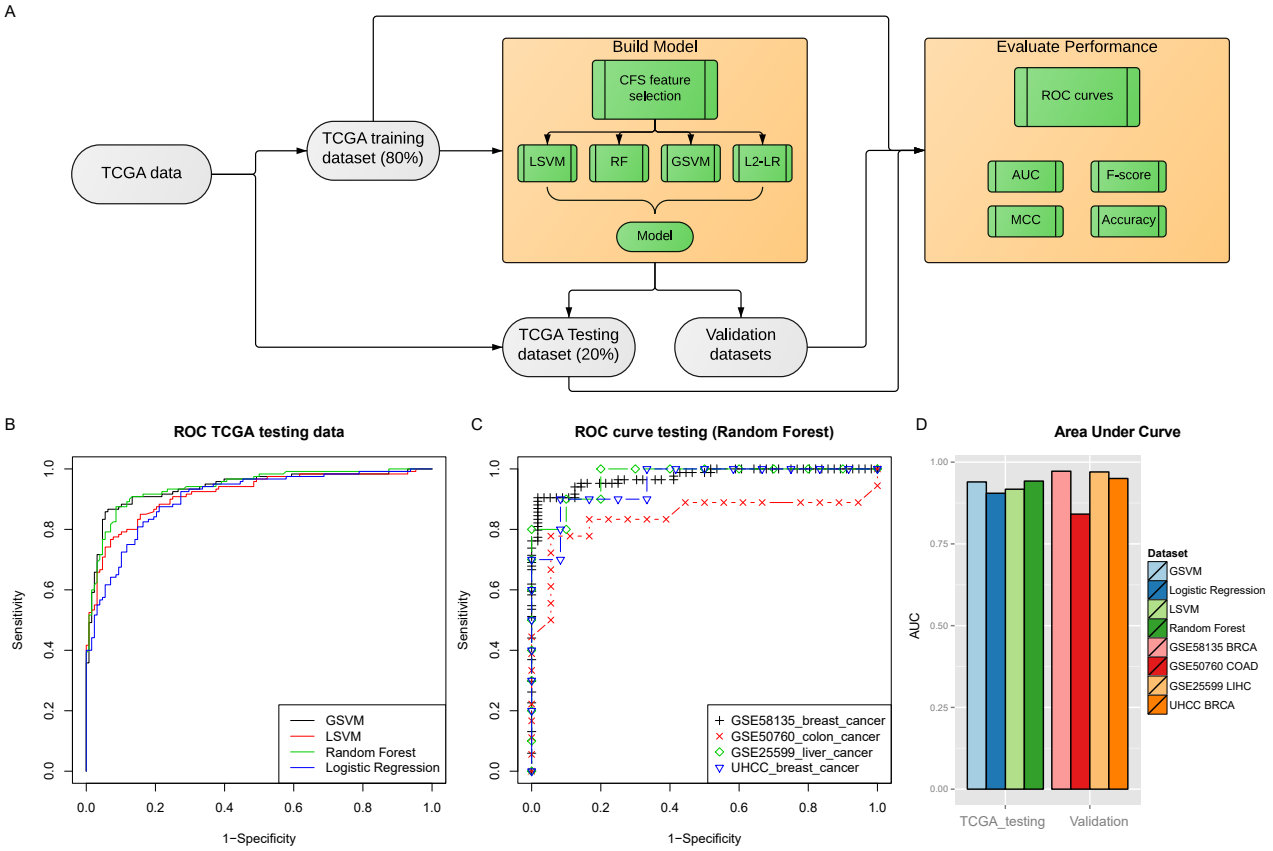


Figure 5: The pan-cancer diagnostic model for the lincRNA panel. (a) The classification of the lincRNA panel was based on a computational RNA-Seq pipeline. The TCGA data were split into 80% training and 20% testing subsets. Five out of the six lincRNAs were selected as predictive features using Correlation Feature Selection (CFS). Pan-cancer diagnostic models were constructed using four standard classification machine learning methods: Random Forest (RF), Linear Support Vector Machines (LSVM), Gaussian Support Vector Machines (GSVM) and Logistic Regression (L2-LR). The best model was chosen based on various metrics of the Receiver operating characteristic (ROC) curves, including Area Under the Curve (AUC), F-score, Matthew's correlation coefficient (MCC) and Accuracy. (b) The performance of the classifier was analysed with the ROC curves on the TCGA hold-out testing data, based on the four classification methods mentioned above and (c) ROC curves of the top Random Forest model on four independent RNA-Seq validation datasets. (d) AUCs were calculated on the TCGA hold-out testing data in and the four validation datasets.

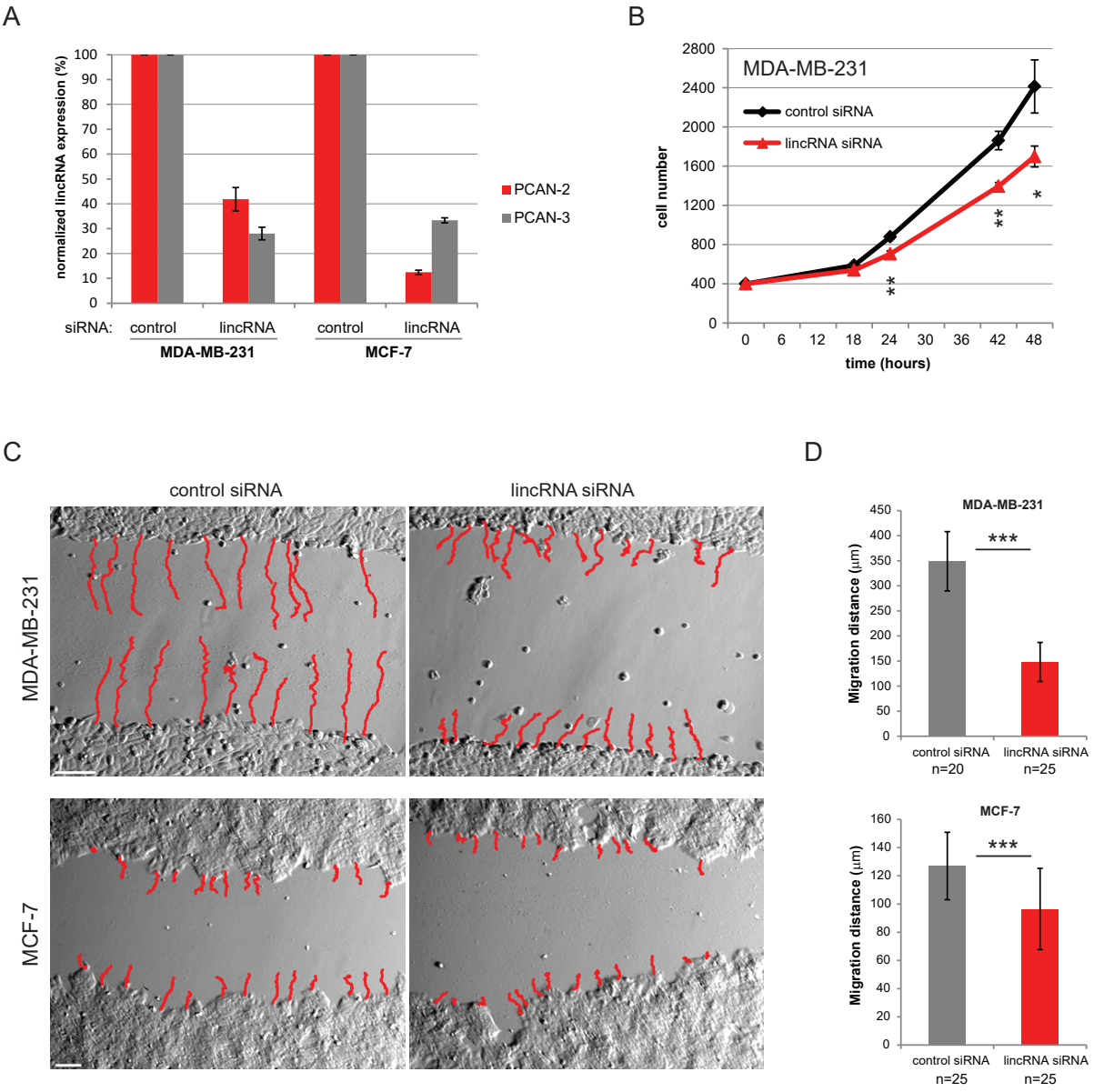


Figure 6: The effect of lincRNAs downregulation on cell proliferation and migration. (a) PCAN-2 and PCAN-3 lincRNAs can be efficiently knocked down in MDA-MB-231 and MCF-7 cell lines. Bars represent RT-qPCR results of the lincRNA expression. siRNA lincRNA bars show mean expression (n=3) with S.D. normalized to the control condition. (b) Transient knockdown of PCAN-2 and PCAN-3 inhibits the growth rate of MDA-MB-231 cells. 30 hours after transfection 400 cells were seeded in 96-well plates and processed for luminescent cell viability assay at indicated time points. Data points represent mean value (n=3), error bars, S.D. *, $P<0.05$, **$P<0.01$. (c) lincRNA knockdown inhibits migration of MDA-MB-231 and MCF-7 cells in wound-healing assay. Cells were transiently transfected 30 hours before making scratches in the cell monolayer. Cell migration rate was analysed with time-lapse microscopy. Red lines – cells tracks analysed over 24 hours. Size bars – 100 micrometers. (d) Quantification of MDA-MB-231 and MCF-7 cells migration distance over 24-hour time period. Bars – value of mean migration distance, error bars +/- S.D. (n=20-25 analysed cells), ***$P<0.001$.

4.9 Appendix

4.9.1 Supplementary figures and tables

Figure S1: Workflow of lincRNA pan-cancer analysis.

TCGA RNA-Seq aligned data (BAM files) and raw data (fastq and SRA files) from several validation datasets were used in this study. Validation datasets were aligned using Tophat2. LincRNAs were quantified using the FeatureCounts program and then normalized as FPKM values. Tissue specificity and subtype analysis was performed on the entire lincRNA transcriptome. Differential expression was analyzed with DESeq2 to obtain a list of pan-cancer lincRNA biomarkers, followed by diagnostic classification modelling and survival analysis.

Figure S2: Principal component analysis of mRNA expression in 12 TCGA datasets.

The first three principal components (PCs) were plotted using the log FPKM values of lincRNA expression in (a) normal adjacent tissue and (b) cancer samples. The variances associated with each of the first 10 principal components are plotted alongside each graph (Scree Plot).

Figure S3: Comparison of lincRNA tissue specificity between TCGA data and Cabili et al1. Each plot is composed of the group of lincRNAs specific to a certain tissue type (liver, kidney etc), as defined in Human Body Map Project by Cabili et al. This group of lincRNAs are reassigned to specific tissues by the JS score calculated from the TCGA data. The correlations between studies in all tissue categories are significant.

Figure S4: JSt scores for each tissue type. Each plot shows the scores for tissue specificity calculated for cancer and normal samples for each individual tissue type.

Figure S5: Differential expression in each cancer type in the 12 TCGA cancer datasets. The significance threshold is set to $\alpha = 0.05 after Benjamini - Hochberg correction.$

Figure S6: Normalized log2 FPKM expression of the panel of six lincRNAs in the 12 TCGA cancer datasets.

Figure S7: Validation of known prognostic lincRNAs. Differential expression analysis of known prognostic lincRNAs markers (PCAT1, MALAT1 and HOTAIR).

Figure S8: Log2 fold change of the six lincRNA panel in the supplementary microarray validation datasets.

Figure S9: LincRNA expression in the breast cancer cell lines from CCLE and GSE58135 compared with primary tumour expression levels.

Figure S10: Correlation heatmap of expression levels among the six lincRNAs.

Figure S11: Blast homology among all transcripts of the six lincRNAs.

Figure S12: ROC of the diagnostic classifier in the TCGA training dataset.

Figure S13: The prognostic potential of the pan-cancer lincRNA panel.

(a-d) The performance of the lincRNA panel in predicting survival is plotted with Kaplan-Meier curves. There were significant differences in overall survival for 463 BRCA patients (a), 350 OV patients with Grade 3 tumours, the dominant grade of TCGA OV (b), as well as the relapse free survival for 193 LUAD patients (c) and 139 LUSC patients (d). The higher and lower risk groups are separated by high and low prognostic index (PI) categories. The PI score is based on the Cox-Regression model of the six lincRNA panel.

Figure 14: Additional cell line experiments on MDA-231 cell line, using siRNA #2 (less efficient siRNA).

(a) lincRNA knockdown by siRNA #2 inhibits migration of MDA-MB-231 in wound-healing assay. Cells were transiently transfected 30 hours before making scratches in the cell monolayer. Cell migration rate was analysed with time-lapse microscopy. Red lines – cells tracks analysed over 24 hours. Size bars – 100 micrometers. (b) Quantification of MDA-MB-231 migration distance over 24-hour time period. Bars – value of mean migration distance, error bars +/- S.D. (n=60 analysed cells), ***P<0.0001.

Figure S15: Additional cell line experiments on HCT116 colon cancer cell line.

(a) PCAN-2 and PCAN-3 lincRNAs can be efficiently knocked down in HCT116 cell lines. Bars represent RT-qPCR results of PCAN-2 (gray) and PCAN-3 (orange) expression. siRNA lincRNA bars show mean expression (n=3) normalized to GUS. (b) Transient knockdown of PCAN-2 and PCAN-3 inhibits the growth rate of HCT-116 cells. 72 hours after transfection. 400 cells were seeded in 96-well plates and processed for luminescent cell viability assay at indicated time points. Data points represent mean value (n=3), error bars, S.D. *, P<0.05. (c) Quantification of HCT116 cells migration distance over 24-hour time period. Bars – value of mean migration distance, error bars +/- S.D. (n=60 analysed cells), *, P=0.036.

Supplementary Tables

Table S1: Tabulation of the patients, tumour samples and normal adjacent tissue samples used in this study.

Table S2: A list of tissue specific lincRNAs and their associated tissue type, defined by JS score > 0.75 and expression in at least 5% of the samples.

Table S3: Genomic coordinate descriptions of the six lincRNAs and cross-reference with the Ensemble database.

Table S4: Primer designs for quantitative real-time PCR of the lincRNA panel.

Table S5: Coding potential predictions of all isoforms of the lincRNA panel, using iSeeRNA and Coding Potential Assessment Tool (CPAT). Additional positive controls using protein-coding genes GAPDH and GUS are also listed.

Table S6: Classification performance metrics of the lincRNA diagnostic model across all datasets (TCGA_training, TCGA_testing, GSE58135_breast_cancer, GSE50760_colon_cancer and GSE25599_liver_cancer). Metrics used are Area Under the Curve (AUC), Accuracy, F-score, Matthew's correlation coefficient (MCC), Sensitivity, Specificity and Precision.

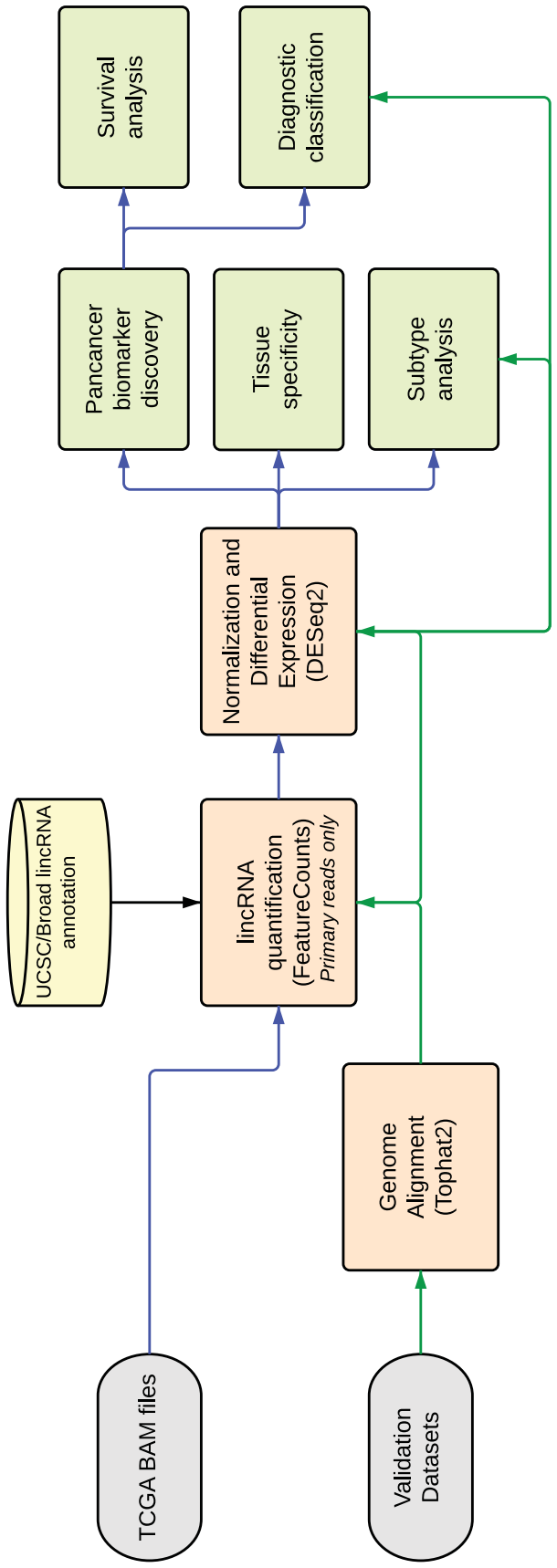

Survival analysis
Diagnostic classification
Pancancer biomarker discovery
Tissue specificity
Subtype analysis
Normalization and Differential Expression (DESeq2)
UCSC/Broad lincRNA annotation
lincRNA quantification (FeatureCounts) Primary reads only
Genome Alignment (Tophat2)
TCGA BAM files
Validation Datasets

Figure S1

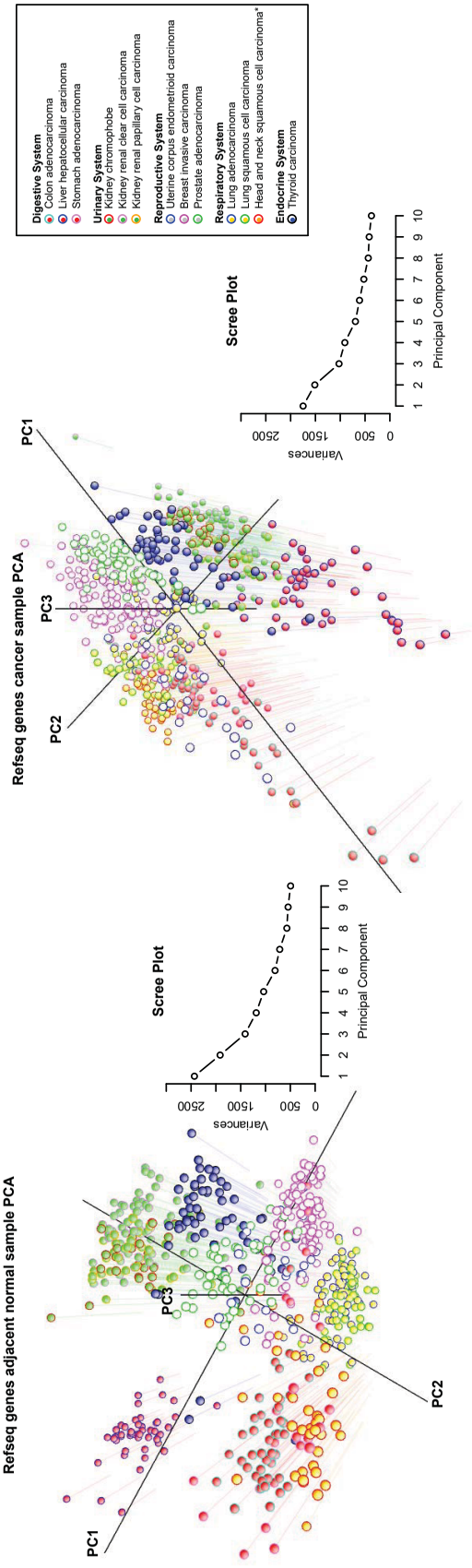

Digestive System
Colon adenocarcinoma
Liver hepatocellular carcinoma
Stomach adenocarcinoma
Urinary System
Kidney chromophobe
Kidney renal clear cell carcinoma
Kidney renal papillary cell carcinoma
Reproductive System
Uterine corpus endometrioid carcinoma
Breast invasive carcinoma
Prostate adenocarcinoma
Respiratory System
Lung adenocarcinoma
Lung squamous cell carcinoma
Head and neck squamous cell carcinoma*
Endocrine System
Thyroid carcinoma
Refseq genes cancer sample PCA
PC1
PC2
PC3
Scree Plot
Variances
Principal Component
Refseq genes adjacent normal sample PCA
PC1
PC2
PC3
Scree Plot
Variances
Principal Component

Figure S2

Figure S3

Histogram of lincRNAs categorized by tissue comparing TCGA lincRNAs versus Cabili et al.

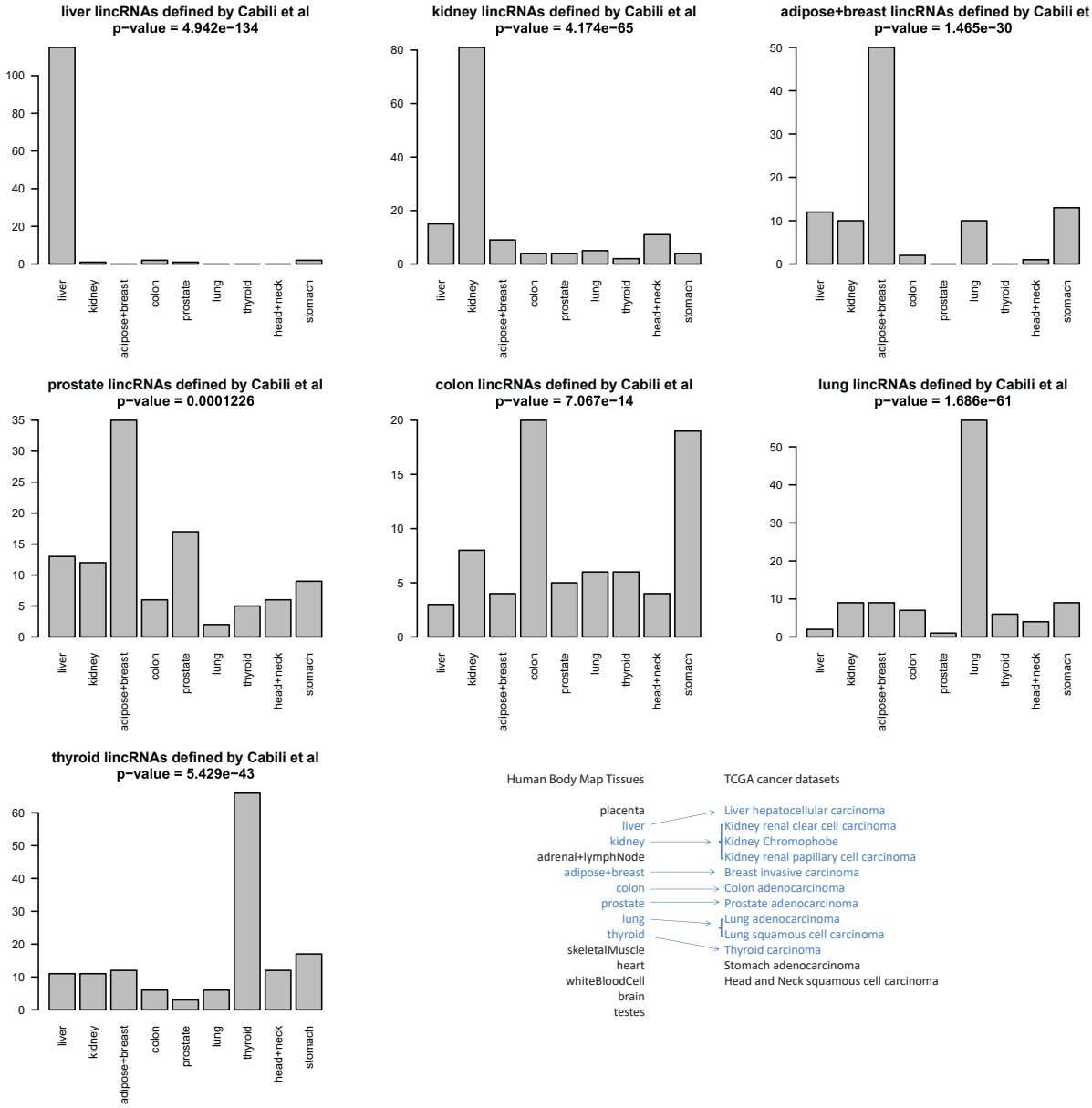


Figure S4

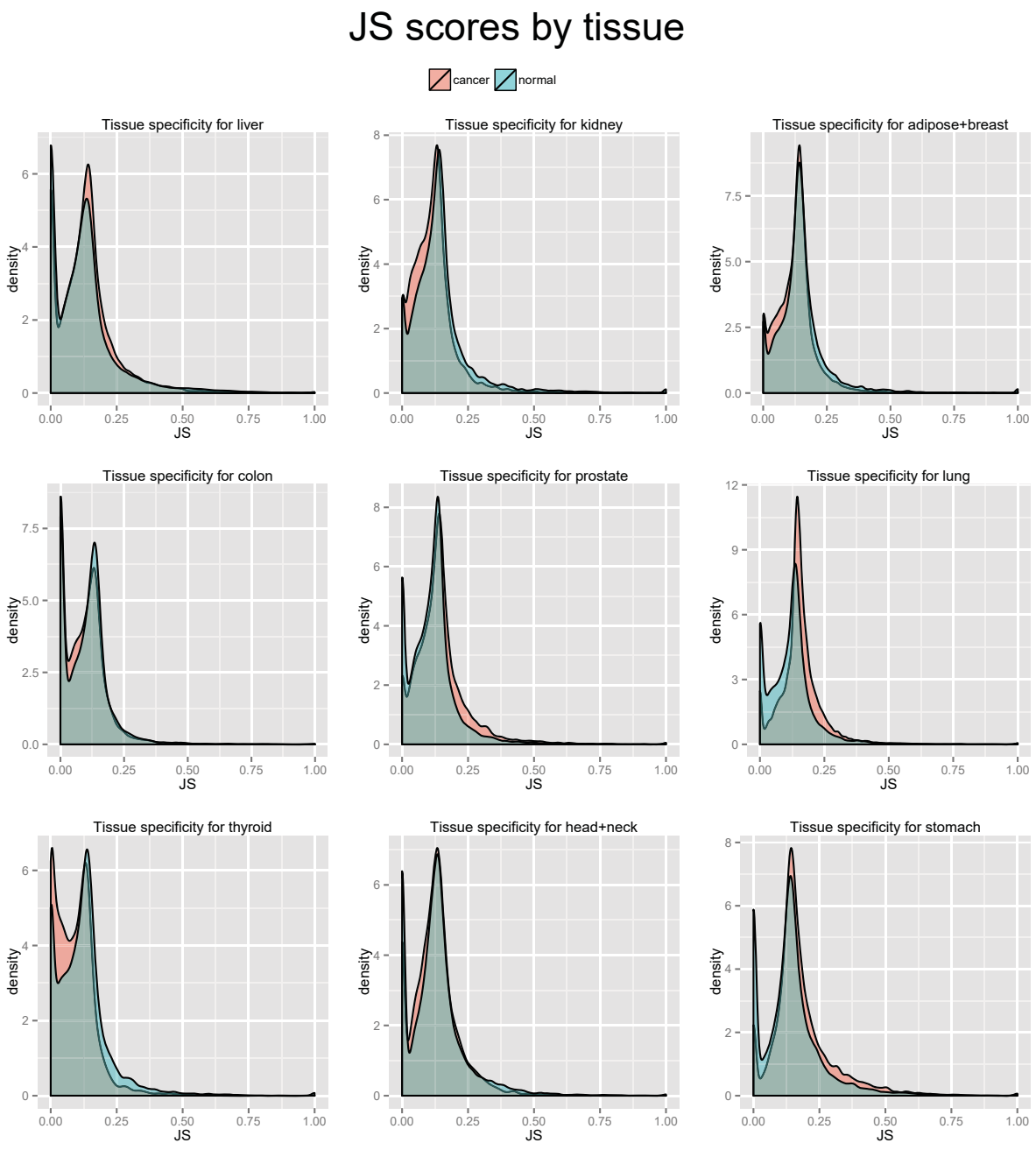

JS scores by tissue
cancer
normal
Tissue specificity for liver
Tissue specificity for kidney
Tissue specificity for adipose+breast
Tissue specificity for colon
Tissue specificity for prostate
Tissue specificity for lung
Tissue specificity for thyroid
Tissue specificity for head+neck
Tissue specificity for stomach
density
JS

Figure S5

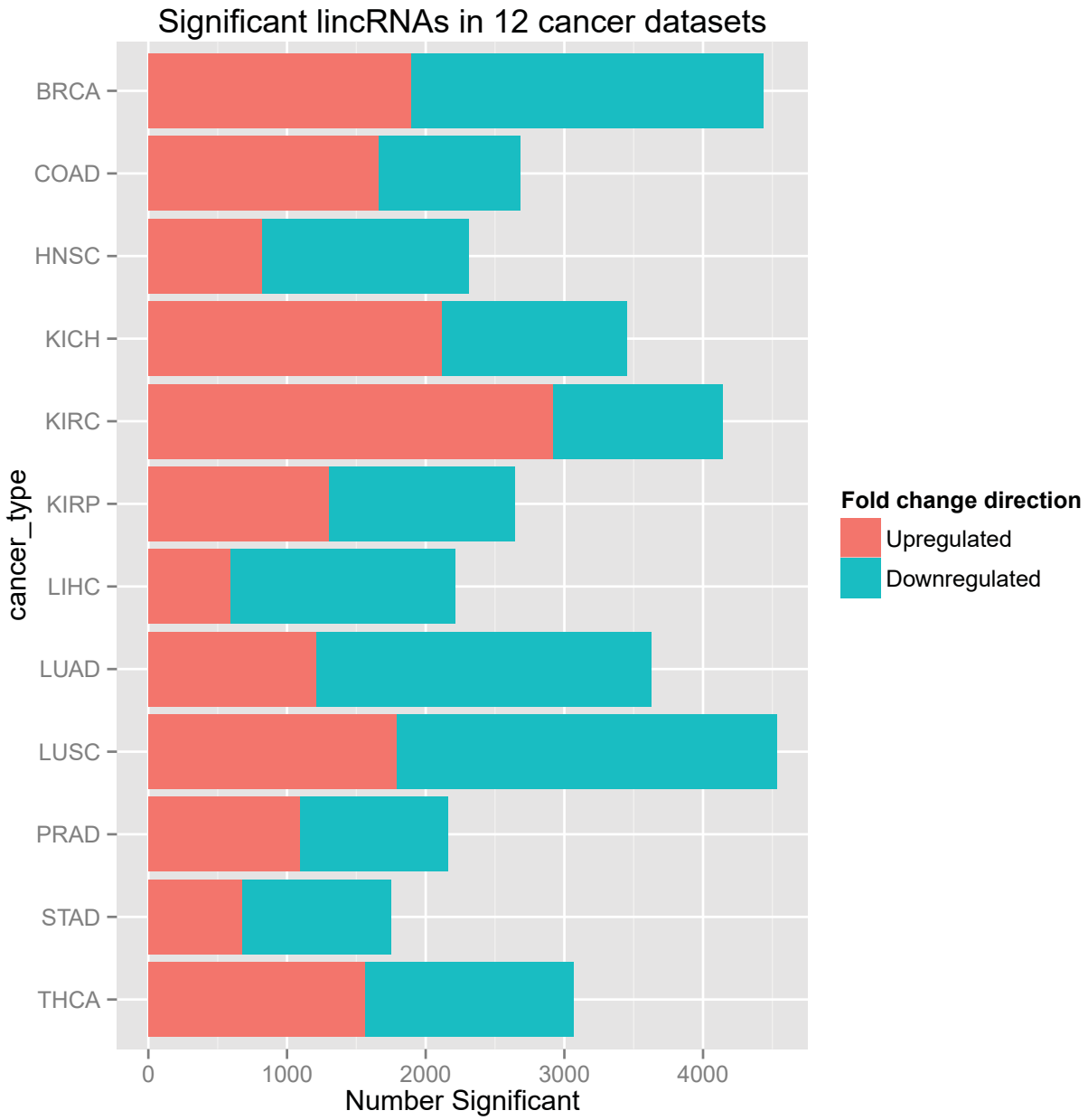

Significant lincRNAs in 12 cancer datasets
BRCA
COAD
HNSC
KICH
KIRC
KIRP
LIHC
LUAD
LUSC
PRAD
STAD
THCA
cancer_type
0
1000
2000
3000
4000
Number Significant
Fold change direction
Upregulated
Downregulated

Figure S6

Figure S7

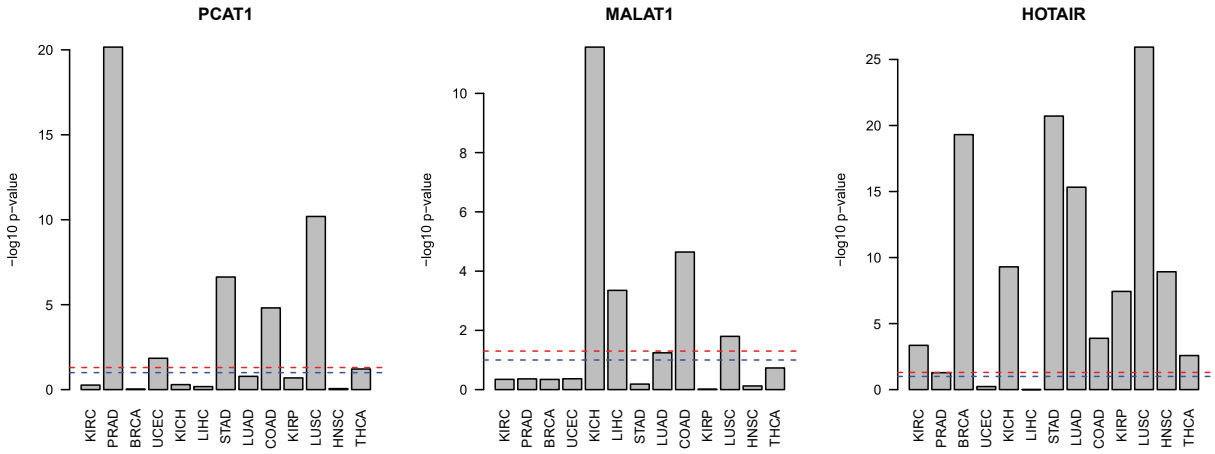


Figure S8

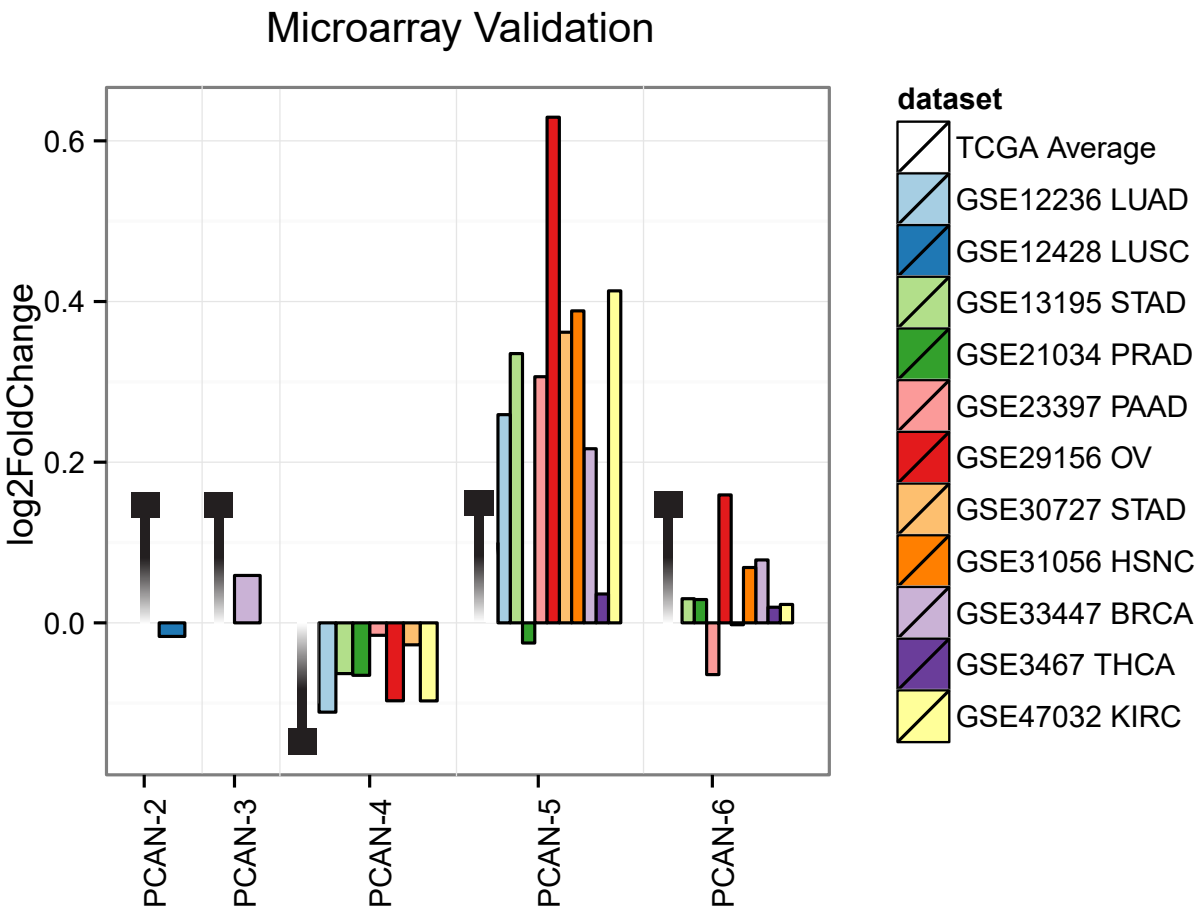


Figure S9

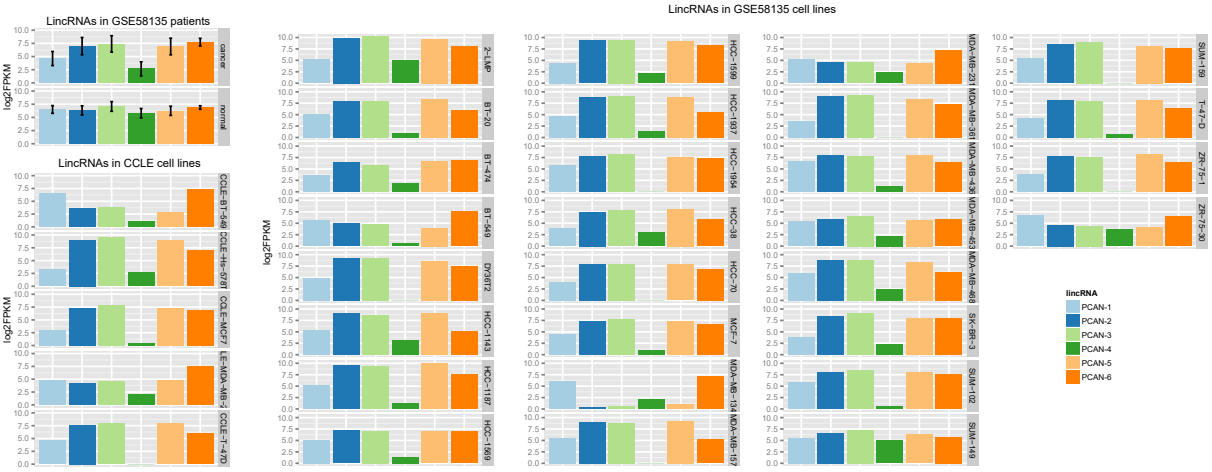


Figure S10

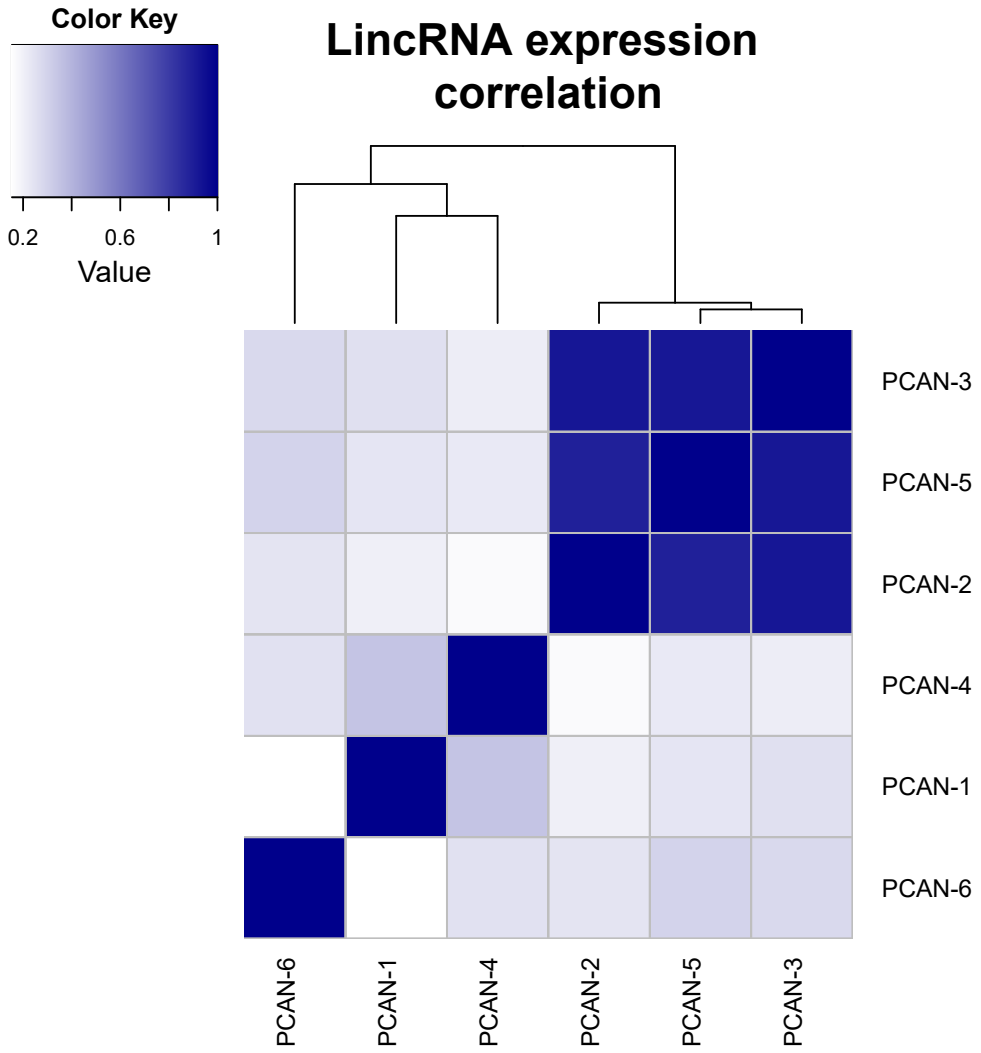

Color Key
0.2
0.6
1
Value
LincRNA expression correlation
PCAN-3
PCAN-5
PCAN-2
PCAN-4
PCAN-1
PCAN-6
PCAN-6
PCAN-1
PCAN-4
PCAN-2
PCAN-5
PCAN-3

Figure S11

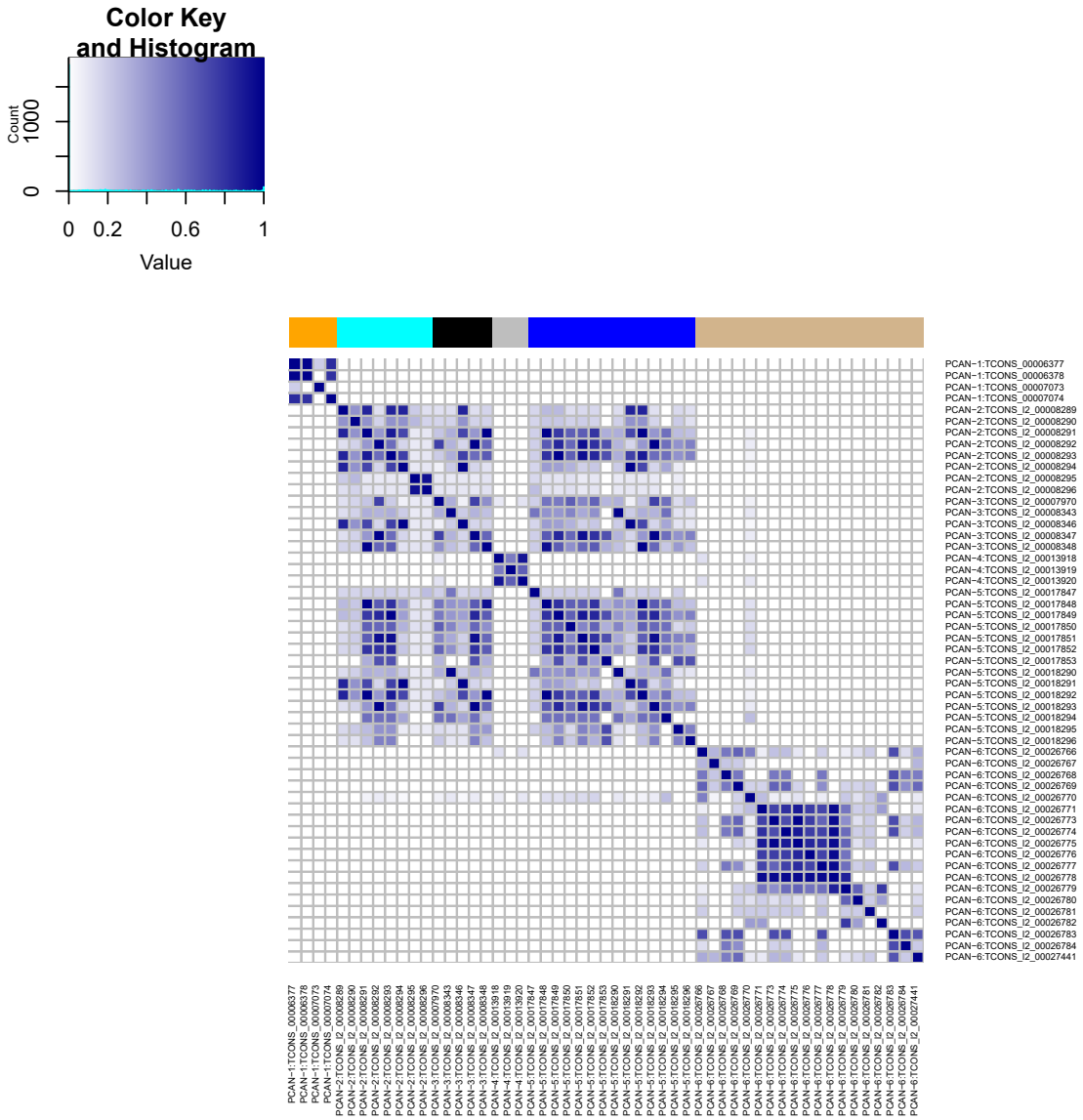

Color Key
and Histogram
Count
1000
0
0
0.2
0.6
1
Value
PCAN−1:TCONS_00006377
PCAN−1:TCONS_00006378
PCAN−1:TCONS_00007073
PCAN−1:TCONS_00007074
PCAN−2:TCONS_I2_00008289
PCAN−2:TCONS_I2_00008290
PCAN−2:TCONS_I2_00008291
PCAN−2:TCONS_I2_00008292
PCAN−2:TCONS_I2_00008293
PCAN−2:TCONS_I2_00008294
PCAN−2:TCONS_I2_00008295
PCAN−2:TCONS_I2_00008296
PCAN−3:TCONS_I2_00007970
PCAN−3:TCONS_I2_00008343
PCAN−3:TCONS_I2_00008346
PCAN−3:TCONS_I2_00008347
PCAN−3:TCONS_I2_00008348
PCAN−4:TCONS_I2_00013918
PCAN−4:TCONS_I2_00013919
PCAN−4:TCONS_I2_00013920
PCAN−5:TCONS_I2_00017847
PCAN−5:TCONS_I2_00017848
PCAN−5:TCONS_I2_00017849
PCAN−5:TCONS_I2_00017850
PCAN−5:TCONS_I2_00017851
PCAN−5:TCONS_I2_00017852
PCAN−5:TCONS_I2_00017853
PCAN−5:TCONS_I2_00018290
PCAN−5:TCONS_I2_00018291
PCAN−5:TCONS_I2_00018292
PCAN−5:TCONS_I2_00018293
PCAN−5:TCONS_I2_00018294
PCAN−5:TCONS_I2_00018295
PCAN−5:TCONS_I2_00018296
PCAN−6:TCONS_I2_00026766
PCAN−6:TCONS_I2_00026767
PCAN−6:TCONS_I2_00026768
PCAN−6:TCONS_I2_00026769
PCAN−6:TCONS_I2_00026770
PCAN−6:TCONS_I2_00026771
PCAN−6:TCONS_I2_00026773
PCAN−6:TCONS_I2_00026774
PCAN−6:TCONS_I2_00026775
PCAN−6:TCONS_I2_00026776
PCAN−6:TCONS_I2_00026777
PCAN−6:TCONS_I2_00026778
PCAN−6:TCONS_I2_00026779
PCAN−6:TCONS_I2_00026780
PCAN−6:TCONS_I2_00026781
PCAN−6:TCONS_I2_00026782
PCAN−6:TCONS_I2_00026783
PCAN−6:TCONS_I2_00026784
PCAN−6:TCONS_I2_00027441

Figure S12

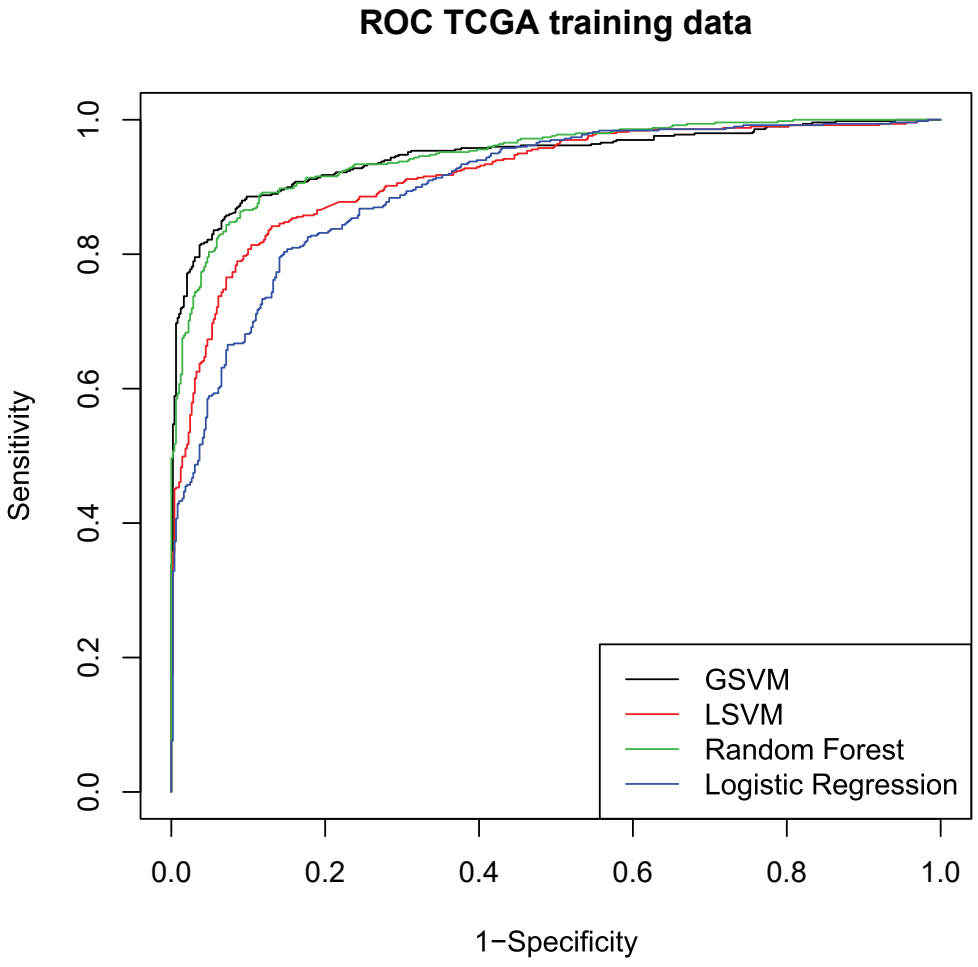

ROC TCGA training data
Sensitivity
1−Specificity
0.0
0.2
0.4
0.6
0.8
1.0
GSVM
LSVM
Random Forest
Logistic Regression

Figure S13

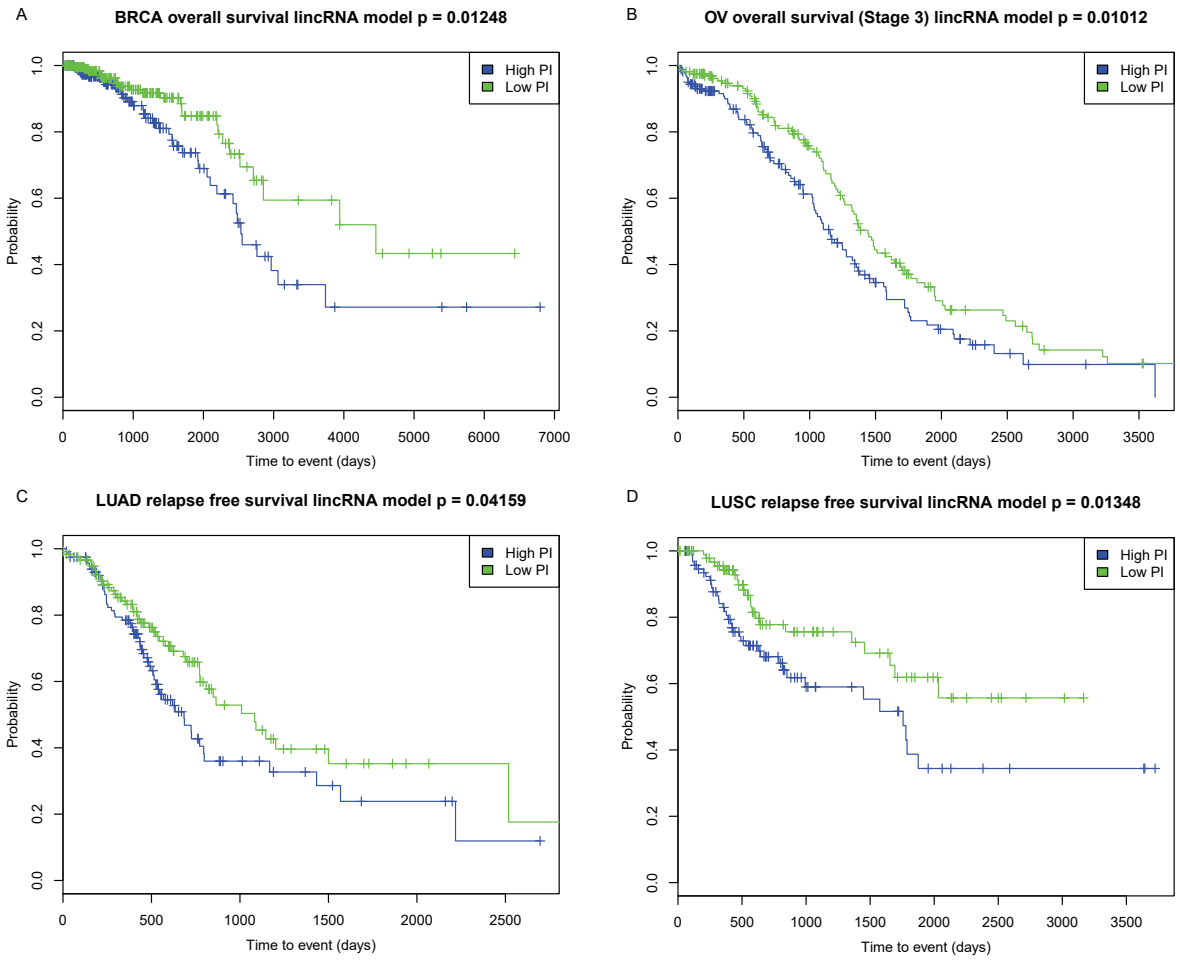

A
BRCA overall survival lincRNA model p = 0.01248
B
OV overall survival (Stage 3) lincRNA model p = 0.01012
C
LUAD relapse free survival lincRNA model p = 0.04159
D
LUSC relapse free survival lincRNA model p = 0.01348
High PI
Low PI
Probability
Time to event (days)

Figure S14

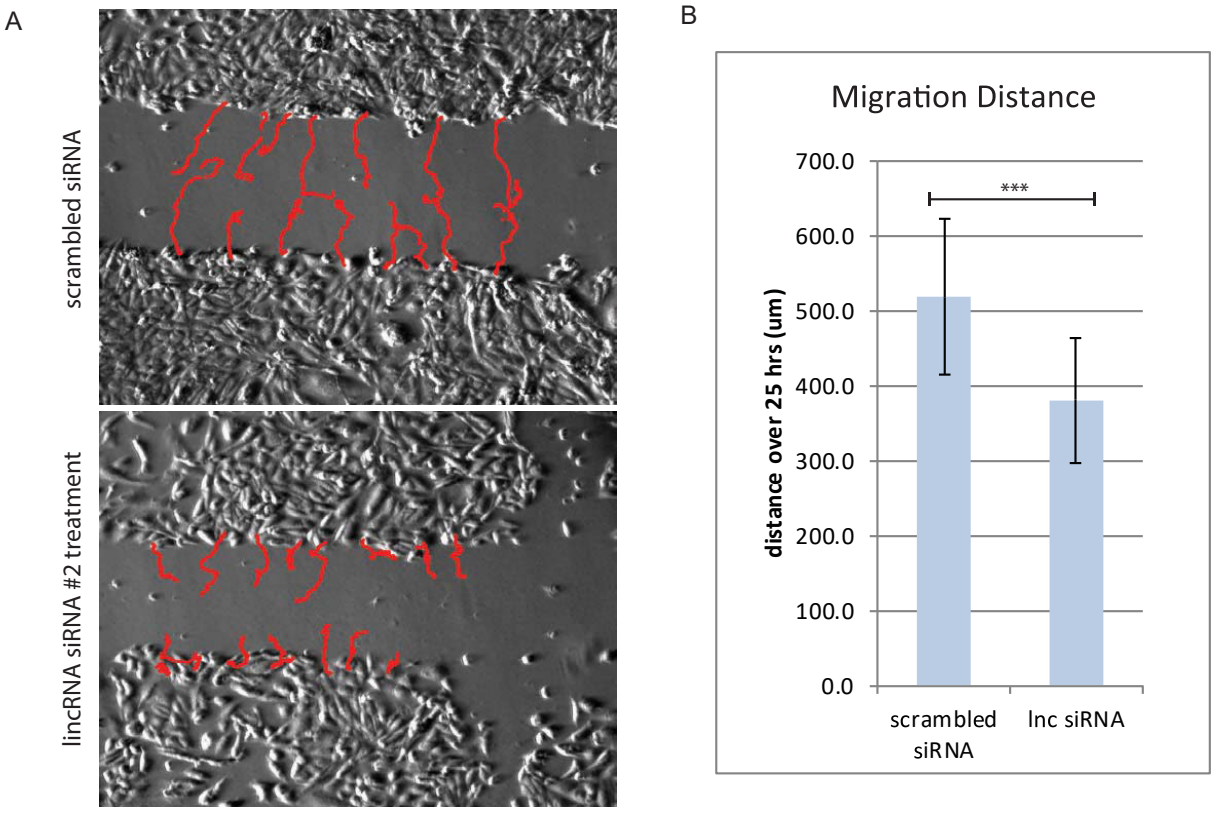
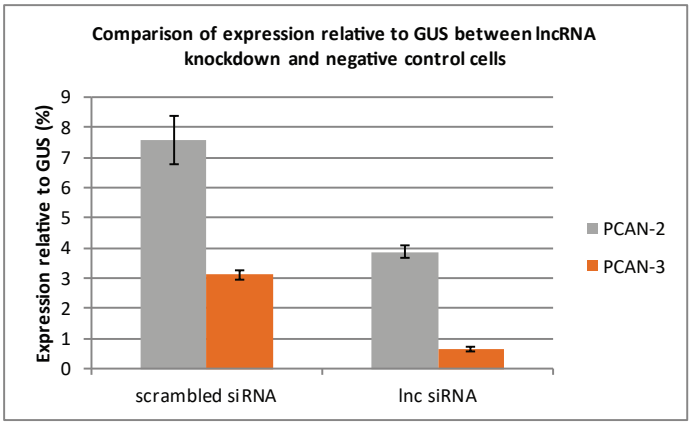
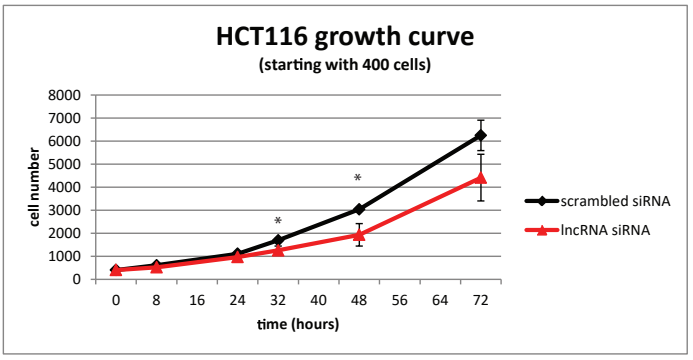

A
scrambled siRNA
lincRNA siRNA #2 treatment
B
Migration Distance
700.0
600.0
500.0
400.0
300.0
200.0
100.0
0.0
***
distance over 25 hrs (um)
scrambled siRNA
lnc siRNA

Figure S15

A



B



C

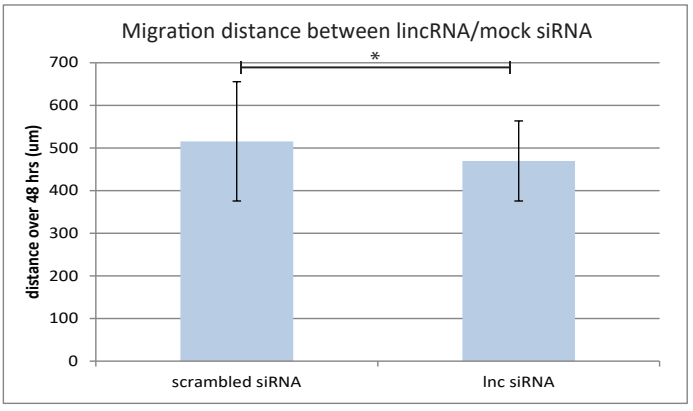


Table S1

TCGA datasets					
Dataset	Primary	Adjacent	Paired	# mapped reads	
Breast invasive carcinoma	1059	111	111	193560468	
Colon adenocarcinoma	444	41	41	44872617	
Head and Neck squamous cell carcinoma	498	43	41	57516555	
Kidney Chromophobe	66	25	25	42982445	
Kidney renal clear cell carcinoma	531	72	72	94701140	
Kidney renal papillary cell carcinoma	226	32	32	48960010	
Liver hepatocellular carcinoma	210	50	50	46819687	
Lung adenocarcinoma	489	58	57	69621821	
Lung squamous cell carcinoma	489	50	50	86702630	
Prostate adenocarcinoma	419	52	52	98002347	
Stomach adenocarcinoma	285	32	30	65146435	
Thyroid carcinoma	498	59	59	127022122	
Subtotal	5214	625	620	975908277	
Samples used for survival analysis	Cancer				
Breast invasive carcinoma	463				
Ovarian serous adenocarcinoma	406				
Lung adenocarcinoma	193				
Lung squamous cell carcinoma	139				
Samples used for subtype analysis	Cancer				
Breast invasive carcinoma	521				
Validation Datasets					
Accssion	Cancer	Non-cancer			Additional Details
GSE25599	10	10			Single end, 36 bp, Liver Cancer (HBV)
GSE58135	84	56			Paired, 100 bp, Breast Cancer
GSE50760	36	36			Paired, 100 bp, Colon cancer
Our BRCA dataset	10	6			Single end, 100 bp, BRCA; normal samples: GSE52194, GSE45326, GSE30611
Whole Human Genome Oligo Microarray G4112A	Cancer	Non-cancer			Additional Details
GSE12428	34	28			Microarray, lung squamous cell carcionoma
Affymetrix Human Exon 1.0 ST Array	Cancer	Non-cancer			Additional Details
GSE30727	30	30			Microarray, stomach cancer
GSE47032	20	20			Microarray, kidney clear cell cancer
GSE23397	15	6			Microarray, pancreatic cancer
GSE21034	150	29			Microarray, prostate cancer
GSE13195	25	25			Microarray, stomach cancer
GSE12236	16	24			Microarray, lung cancer
GSE29156	9	11			Microarray, ovarian cancer
Affymetrix Human Genome U133 Plus 2.0 Array	Cancer	Non-cancer			Additional Details
GSE3467	9	9			Microarray, thyroid samples
GSE31056	23	72			Microarray, oral carcinoma
SurePrint G3 Human GE 8x60K Microarray	Cancer	Non-cancer			Additional Details
GSE33447	8	8			Microarray, breast cancer samples
	ER+/HER2-	Triple Negative			
GSE58135 subtype analysis	42	42			
Total unique biological samples	3354				

Table S2

lincRNA	JS_Score	Associated_Tissue
XLOC_010690	0.766204217	kidney
XLOC_007596	0.857703592	liver
XLOC_003732	0.752100476	colon
XLOC_010399	0.768482652	kidney
XLOC_008645	0.758646776	lung
XLOC_011257	0.763064551	liver
XLOC_001387	0.77810803	liver
XLOC_007597	0.824089647	liver
XLOC_000947	0.903394197	adipose+breast
XLOC_011275	0.911613639	liver
XLOC_004719	0.941948946	kidney
XLOC_013795	0.781746482	adipose+breast
XLOC_004102	0.873732325	kidney
XLOC_004360	0.785572253	adipose+breast
XLOC_003239	0.775039257	lung
XLOC_008455	0.770565259	adipose+breast
XLOC_004177	1	lung
XLOC_001901	0.760514635	liver
XLOC_004836	0.754396243	adipose+breast
XLOC_004514	0.794355439	liver
XLOC_004801	0.906048522	adipose+breast
XLOC_004261	0.817042521	adipose+breast
XLOC_000224	0.799618236	kidney
XLOC_004269	0.867461193	kidney
XLOC_008454	0.792515074	adipose+breast
XLOC_004515	0.752963989	liver
XLOC_005515	0.767317503	thyroid
XLOC_008260	0.902906732	lung
XLOC_001704	0.848966352	liver
XLOC_013779	0.756769238	head+neck
XLOC_007883	0.926395141	liver
XLOC_006805	0.772086177	kidney
XLOC_008446	1	kidney
XLOC_003315	0.791869141	colon
XLOC_012693	0.755588854	liver
XLOC_000857	0.760980892	kidney
XLOC_009690	0.763803004	lung
XLOC_001425	0.900101922	kidney
XLOC_005607	0.808412113	liver
XLOC_010419	0.779716586	thyroid
XLOC_013465	0.754345459	prostate
XLOC_007017	0.773648455	adipose+breast
XLOC_008602	0.767872806	liver
XLOC_007405	0.804291982	prostate
XLOC_014355	0.762921632	prostate
XLOC_013037	0.768622905	prostate
XLOC_001378	0.862077396	kidney
XLOC_005775	0.817266832	kidney
XLOC_000430	0.936023257	liver
XLOC_l2_000357	0.767464299	lung
XLOC_l2_013883	0.914582666	prostate
XLOC_l2_001548	0.763547727	adipose+breast
XLOC_l2_004342	0.768931463	prostate

Table S3

lincRNA	Human Body Map ID	Chromosome	lincRNA strand	lincRNA start	lincRNA width	lincRNA end
PCAN-1	XLOC_002996	chr3	+	195367106	10516	195377622
PCAN-2	XLOC_l2_004121	chr14	+	19650018	45163	19695181
PCAN-3	XLOC_l2_004340	chr14	-	19856361	68973	19925334
PCAN-4	XLOC_l2_007509	chr2	+	114298969	29137	114328106
PCAN-5	XLOC_l2_009441	chr22	-	16101370	91857	16193227
PCAN-6	XLOC_l2_013931	chr7	-	97503667	98000	97601667

Closest Gene	Distance	Gene start	Gene width	Gene end	Ensembl gene match	Ensembl Annotation (GRCh37)
APOD	56029	195295573	15504	195311077	N/A	LincRNA, Unannotated
POTEG	65075	19553365	31578	19584943	AL589743.1	LincRNA, Unannotated
POTEM	58620	19983954	36319	20020273	CTD-2314B22.3	LincRNA, Unannotated
FOXD4L1	40241	114256661	2067	114258728	N/A	Processed Pseudogene, Unannotated
POTEH	63105	16256332	31606	16287938	AP000525.9	LincRNA, Processed Transcript, Unannotated
ASNS	1812	97481429	20426	97501855	N/A	Processed Pseudogene, Unannotated

Table S4

Assay Label	LincRNA target	Primers
4121_1:	PCAN-2 (XLOC_l2_004121)	F: 5'-AGCTTCGGAGAAGCAGTGGT-3'
		R: 5'-TTCTTTCCGCGGAGACCT-3'
4340_4	PCAN-3 (XLOC_l2_004340)	F: 5'-ACAGATGAACCGCGGAGAC-3'
		R: 5'-AGCTTCGGAGAAGCAGTGGT-3'
2996_3	PCAN-1 (XLOC_l2_002996)	F: 5'-TAAGGGTCATGGAGCTGGAG-3'
		R: 5'-ATCAGCTCCTCCCCGAGTAT-3'
7509_4	PCAN-4 (XLOC_l2_007509)	F: 5'-GAAGTTTAATGTTGCCAATGGA-3'
		R: 5'-GCCTTTGCACAGACTGACCT-3'
13931_6	PCAN-6 (XLOC_l2_013931)	F: 5'-ATCCAGAACTGCAGCCAGTC-3'
		R: 5'-AGAAGTACATGGGGGTGTGG-3'

Table S5 part 1

http://lilab.research.bcm.edu/cpat/calculator_sub.php
Reference: CPAT: Coding-Potential Assessment Tool using an alignment-free logistic regression model. Nucleic Acid Research

PCAN-1 (XLOC_002996)

Sequence Name	RNA Size	ORF Size	Ficket Score	Hexamer Score	Coding Probability	Coding Label
TCONS_00006377	6659	633	0.5291	0.173947044	0.760142576	yes
TCONS_00006378	1716	135	1.0111	0.171563902	0.080769211	no
TCONS_00007073	211	96	0.6124	-0.376596856	0.000483433	no
TCONS_00007074	1586	153	1.1382	0.213026381	0.17417223	no

PCAN-2 (XLOC_l2_004121)

Sequence Name	RNA Size	ORF Size	Ficket Score	Hexamer Score	Coding Probability	Coding Label
TCONS_L2_00008289	768	117	0.7464	0.041918727	0.013971842	no
TCONS_L2_00008290	549	192	0.8298	0.048534387	0.042017289	no
TCONS_L2_00008291	2102	276	0.484	-0.012649119	0.021810528	no
TCONS_L2_00008292	1406	327	0.639	-0.180683866	0.021204847	no
TCONS_L2_00008293	2391	267	0.42	-0.077650944	0.010473072	no
TCONS_L2_00008294	4116	267	0.42	-0.077650944	0.009441453	no
TCONS_L2_00008295	1060	342	0.5027	-0.119244657	0.02480705	no
TCONS_L2_00008296	990	342	0.5027	-0.119244657	0.024910079	no

PCAN-3 (XLOC_l2_004340)

Sequence Name	RNA Size	ORF Size	Ficket Score	Hexamer Score	Coding Probability	Coding Label
TCONS_L2_00007970	5950	459	0.6599	0.112352222	0.336164826	no
TCONS_L2_00008343	1169	144	0.6051	-0.107657632	0.004404995	no
TCONS_L2_00008346	4142	267	0.42	-0.077650944	0.009426701	no
TCONS_L2_00008347	1405	327	0.639	-0.180683866	0.021206108	no
TCONS_L2_00008348	570	168	0.904	-0.068320034	0.019342915	no

PCAN-4 (XLOC_l2_007509)

Sequence Name	RNA Size	ORF Size	Ficket Score	Hexamer Score	Coding Probability	Coding Label
TCONS_L2_00013918	2879	270	0.6819	-0.343879747	0.004149952	no
TCONS_L2_00013919	547	192	1.0539	0.129150962	0.131216094	no
TCONS_L2_00013920	2179	270	0.6819	-0.343879747	0.00432936	no

PCAN-5 (XLOC_l2_009441)

Sequence Name	RNA Size	ORF Size	Ficket Score	Hexamer Score	Coding Probability	Coding Label
TCONS_L2_00017847	609	150	0.8864	-0.136597695	0.009670352	no
TCONS_L2_00017848	880	168	0.868	-0.07522081	0.016248197	no
TCONS_L2_00017849	913	168	0.868	-0.07522081	0.016216201	no
TCONS_L2_00017850	763	168	0.868	-0.07522081	0.016362141	no
TCONS_L2_00017851	757	144	0.6203	-0.079415711	0.005701195	no
TCONS_L2_00017852	631	168	0.868	-0.07522081	0.016491637	no
TCONS_L2_00017853	359	93	0.9437	-0.087655495	0.008776281	no
TCONS_L2_00018290	1174	144	0.6051	-0.107657632	0.004403664	no
TCONS_L2_00018291	4136	267	0.42	-0.077650944	0.009430103	no
TCONS_L2_00018292	2102	276	0.484	-0.012649119	0.021810528	no
TCONS_L2_00018293	1394	327	0.6533	-0.197474682	0.019892986	no
TCONS_L2_00018294	413	144	0.6051	-0.107657632	0.004610957	no
TCONS_L2_00018295	871	102	0.8213	-0.089455015	0.006323578	no
TCONS_L2_00018296	551	258	0.6655	0.081717204	0.062360503	no

PCAN-6 (XLOC_l2_013931)

Sequence Name	RNA Size	ORF Size	Ficket Score	Hexamer Score	Coding Probability	Coding Label
TCONS_L2_00026766	1676	459	0.484	0.100992688	0.259495691	no
TCONS_L2_00026767	809	591	0.702	0.137974424	0.795810077	yes
TCONS_L2_00026768	376	252	1.047	0.176429194	0.280760562	no
TCONS_L2_00026769	453	153	1.0189	0.316429065	0.23511127	no
TCONS_L2_00026770	4110	396	0.618	-0.361689179	0.010864527	no
TCONS_L2_00026771	1215	267	1.1441	0.363689952	0.67084369	yes
TCONS_L2_00026773	1150	267	1.1441	0.363689952	0.671714583	yes
TCONS_L2_00026774	1091	267	1.1441	0.363689952	0.672504065	yes
TCONS_L2_00026775	829	267	1.1441	0.363689952	0.675998075	yes
TCONS_L2_00026776	930	327	1.2265	0.349269552	0.823711195	yes
TCONS_L2_00026777	1119	267	1.1441	0.363689952	0.672129517	yes
TCONS_L2_00026778	582	267	1.1441	0.363689952	0.679274184	yes
TCONS_L2_00026779	1167	267	1.1441	0.363689952	0.671486925	yes
TCONS_L2_00026780	659	135	0.3685	-0.40494765	0.00027646	no
TCONS_L2_00026781	400	156	0.6087	-0.431469675	0.000628962	no
TCONS_L2_00026782	255	246	0.4729	-0.16219555	0.006446424	no
TCONS_L2_00026783	287	204	0.9136	0.316628344	0.279341985	no
TCONS_L2_00026784	1549	210	1.0844	0.305066245	0.377734015	yes
TCONS_L2_00027441	929	426	0.4914	0.202855179	0.339790484	no

Protein coding gene controls

GAPDH

Sequence Name	RNA Size	ORF Size	Ficket Score	Hexamer Score	Coding Probability	Coding Label
ENST00000229239.5	1875	1008	1.2926	0.519960341	0.999962146	yes
ENST00000396861.1	1348	1008	1.2926	0.519960341	0.999963338	yes
ENST00000396858.1	1292	882	1.2952	0.533919114	0.999870952	yes
ENST00000396859.1	1256	1008	1.2926	0.519960341	0.999963542	yes
ENST00000396856.1	1266	783	1.315	0.548545855	0.999679465	yes

GUSB

Sequence Name	RNA Size	ORF Size	Ficket Score	Hexamer Score	Coding Probability	Coding Label
ENST00000304895.4	2300	339	1.2668	0.277977644	0.776289336	yes
ENST00000421103.1	1742	339	1.2668	0.277977644	0.782118123	yes
ENST00000345660.6	2027	285	1.1446	0.333726267	0.659495687	yes

Table S5 part 2

Results from iSeeRNA: http://137.189.133.71/iSeeRNA/

PCNA-1 (XLOC_002996)

ID	C/NC	NONCODING SCORE
TCONS_00007074	noncoding	0.9837
TCONS_00007073	noncoding	0.996
TCONS_00006377	noncoding	0.9873
TCONS_00006378	noncoding	0.9828

PCAN-2 (XLOC_l2_004121)

ID	C/NC	NONCODING SCORE
TCONS_l2_00008295	noncoding	0.8142
TCONS_l2_00008296	noncoding	0.7723
TCONS_l2_00008290	noncoding	0.9384
TCONS_l2_00008289	noncoding	0.9267
TCONS_l2_00008292	noncoding	0.9812
TCONS_l2_00008294	noncoding	0.9769
TCONS_l2_00008291	noncoding	0.9582
TCONS_l2_00008293	noncoding	0.9681

PCAN-3 (XLOC_l2_004340)

ID	C/NC	NONCODING SCORE
TCONS_l2_00008348	noncoding	0.9764
TCONS_l2_00008347	noncoding	0.9819
TCONS_l2_00008343	noncoding	0.8167
TCONS_l2_00008346	noncoding	0.9733
TCONS_l2_00007970	noncoding	0.9729

PCAN-4 (XLOC_l2_007509)

ID	C/NC	NONCODING SCORE
TCONS_l2_00013919	coding	0.208638
TCONS_l2_00013920	noncoding	0.9405
TCONS_l2_00013918	noncoding	0.6351

PCAN-5 (XLOC_l2_009441)

ID	C/NC	NONCODING SCORE
TCONS_l2_00017847	noncoding	0.6722
TCONS_l2_00017848	noncoding	0.9813
TCONS_l2_00017849	noncoding	0.9852
TCONS_l2_00017850	noncoding	0.9341
TCONS_l2_00017851	noncoding	0.9674
TCONS_l2_00017852	noncoding	0.8795
TCONS_l2_00017853	noncoding	0.984
TCONS_l2_00018290	noncoding	0.8125
TCONS_l2_00018291	noncoding	0.9676
TCONS_l2_00018292	noncoding	0.9702
TCONS_l2_00018293	noncoding	0.9803
TCONS_l2_00018294	noncoding	0.9135
TCONS_l2_00018295	noncoding	0.9988
TCONS_l2_00018296	noncoding	0.9808

PCAN-6 (XLOC_l2_013931)

ID	C/NC	NONCODING SCORE
TCONS_l2_00026774	noncoding	0.6841
TCONS_l2_00026769	coding	0.0722
TCONS_l2_00026766	coding	0.2569
TCONS_l2_00026779	noncoding	0.8778
TCONS_l2_00026780	noncoding	0.8494
TCONS_l2_00026781	noncoding	0.9237
TCONS_l2_00026767	noncoding	0.7907
TCONS_l2_00026771	noncoding	0.872
TCONS_l2_00026778	noncoding	0.6522
TCONS_l2_00026776	noncoding	0.7908
TCONS_l2_00026777	noncoding	0.605
TCONS_l2_00027441	coding	0.141
TCONS_l2_00026784	noncoding	0.8035
TCONS_l2_00026768	noncoding	0.8508
TCONS_l2_00026773	noncoding	0.8681
TCONS_l2_00026770	noncoding	0.6034
TCONS_l2_00026783	noncoding	0.7266
TCONS_l2_00026782	noncoding	0.5388
TCONS_l2_00026775	noncoding	0.8405

Protein coding gene controls

GAPDH

ID	C/NC	NONCODING SCORE
ENST00000229239.5	coding	0
ENST00000396861.1	coding	0
ENST00000396858.1	coding	0.002
ENST00000396859.1	coding	0
ENST00000396856.1	coding	0.0038

GUSB

ID	C/NC	NONCODING SCORE
ENST00000304895.4	coding	0
ENST00000421103.1	coding	0
ENST00000345660.6	coding	0

Table S6

Area under curve						
	TCGA_training	TCGA_testing	GSE58135_breast_cancer	GSE50760_colon_cancer	GSE25599_liver_cancer	Our BRCA dataset
GSVM	0.946	0.939	0.545	0.827	0.910	0.750
LSVM	0.918	0.917	0.968	0.796	0.885	1.000
Random Forest	0.947	0.942	0.972	0.841	0.970	0.950
Logistic Regression	0.901	0.905	0.646	0.410	0.390	0.550
Max accuracy						
	TCGA_training	TCGA_testing	GSE58135_breast_cancer	GSE50760_colon_cancer	GSE25599_liver_cancer	Our BRCA dataset
GSVM	0.894	0.903	0.743	0.861	0.850	0.773
LSVM	0.856	0.851	0.921	0.833	0.850	1.000
Random Forest	0.887	0.895	0.936	0.861	0.900	0.909
Logistic Regression	0.828	0.831	0.679	0.639	0.600	0.636
F-score						
	TCGA_training	TCGA_testing	GSE58135_breast_cancer	GSE50760_colon_cancer	GSE25599_liver_cancer	Our BRCA dataset
GSVM	0.788	0.808	0.489	0.732	0.734	0.624
LSVM	0.712	0.708	0.836	0.684	0.734	1.000
Random Forest	0.775	0.790	0.873	0.732	0.816	0.817
Logistic Regression	0.658	0.665	0.418	0.302	0.250	0.346
Matthew's correlation coeficient						
	TCGA_training	TCGA_testing	GSE58135_breast_cancer	GSE50760_colon_cancer	GSE25599_liver_cancer	Our BRCA dataset
GSVM	0.894	0.897	0.822	0.848	0.870	0.800
LSVM	0.855	0.843	0.936	0.824	0.842	1.000
Random Forest	0.888	0.890	0.944	0.848	0.909	0.900
Logistic Regression	0.826	0.835	0.750	0.667	0.667	0.645

4.10 Chapter summary

In this chapter, I quantify the lincRNA expression in over 3000 unique patient samples from 12 different cancer types, in both primary tumor samples and paired normal adjancent samples. We explore how the expression of lincRNAs relate to clinical phenotypes, such as overall patient survival, disease free sruvival and tumor subtypes. We find that there are six lincRNAs that are consistently differentially expressed.

We show in both the TCGA samples and several validation datasets that these lincRNAs can be used as a novel lincRNA biomarker panel that can accurately distinguish between tumor and normal samples. The expression of these six lincRNAs also seem to be significantly correlated with patient survival. Finally, we show that there is some change of cancer phenotypes through knockdown of the lincRNAs on cancer cell lines.

Chapter 5

Cox-nnet: an artificial neural network method for prognosis prediction of high-throughput omics data

Travers Ching[1,2], Xun Zhu[1,2], Lana X. Garmire[1,2]

Submitted to *PLOS Computational Biology*, under revision.

[1]University of Hawaii Cancer Center, 701 Ilalo Street, Honolulu, Hawaii, USA 96813

[2]Graduate Program of Molecular Biosciences and Bioengineering, University of Hawaii-Manoa, 1955 East-West Road, Honolulu, Hawaii, USA 96822

5.1 Preface

Artificial neural networks (ANN) are computing architectures with many interconnections of simple neural-inspired computing elements, and have been applied to biomedical fields such as imaging analysis and diagnosis. We have developed a new ANN framework called Cox-nnet to predict patient prognosis from high throughput transcriptomics data. In 10 TCGA RNA-Seq data sets, Cox-nnet achieves a statistically significant increase in predictive accuracy compared to other methods, including Cox-proportional hazards regression (with LASSO, ridge,

and mimimax concave penalty), Random Forests Survival and CoxBoost. Cox-nnet also reveals richer biological information, at both the pathway and gene levels. The outputs from the hidden layer node provide an alternative approach for survival-sensitive dimension reduction. In summary, we have developed a new method for more accurate and efficient prognosis prediction on high throughput data, with functional biological insights. The source code is freely available at `https://github.com/lanagarmire/cox-nnet`.

5.2 Author Summary

The increasing application of high-througput transcriptomics data to predict patient prognosis demand modern computational methods. With the re-gaining popularity of artificial neural network, we asked if a refined neural network model could be used to predict patient survival, as an alternative to the conventional methods, such as Cox proportional hazards (Cox-PH) methods with LASSO or ridge penalization. To this end, we have developed a neural network extension of the Cox regression model, called Cox-nnet. It is optimized for survival prediction from high throughput gene expression data. Cox-nnet is statistically more accurate than previous algorithms. Moreover, Cox-nnet reveals much richer biological information, at both the pathway and gene levels, by analyzing features represented in the hidden layer nodes in Cox-nnet. Additionally, we propose to use hidden node features as a new approach for dimension reduction during survival data analysis.

5.3 Introduction

The human brain is a neural network consisting of 10^{11} interconnected neurons [1]. Each neuron receives input from other neurons, and under certain conditions it is activated and outputs its own signals as electric pulses. Inspired by this structure, Artificial Neural networks (ANNs) were developed in 1943 to models non-linear behavior [2]. In ANN, hidden units, termed as neurons or nodes, may be activated or deactivated, depending on the input signals, their own linear weight and bias parameters. The data are fed forward through the network, and for each hidden unit, these weight and bias parameters are learned through backpropagation of the information along the gradient of the loss function. In recent years, ANNs have caught renewed attention, thanks to increased parallel computing power and the promise of deep learning [3].

Survival analysis is a regression problem in order to model patients survival time (or other event time). However, unlike standard regression problems, survival analysis is complicated by

censoring, where the subject leaves the study before the occurrence of the event. To deal with this issue, semi-parametric proportional hazards models were developed, where the covariates of the models explain the relative risks of the patients, termed hazard ratios. The first ANN model to predict survival was done by Faraggi and Simon, who only used four clinical input parameters to model prostate cancer survival[4]. However, their simple model was by no means suitable for high throughput input data. Subsequently, other authors attempted to implement ANN methods to predict patient survival. One study applied ANNs to high dimensional survival data by simplifying the regression as a binary classification problem [5, 6], and another study fit continuous variables of survival time to discrete variables through binning [5, 6]. These approaches potentially led to loss of accuracy in prediction. Another study used time as an additional input in order to predict patient survival or censoring status [7], with the potential to overfit when the survival and censoring are correlated. Thus far, a genuine ANN model based on proportional hazards designed to analyze high throughput data in the genomics era is lacking.

Meanwhile there exist other modeling approaches for patient survival prediction. The most common method is the Cox-PH model, a linear model based on the proportional hazards [8]. In this manuscript, we evaluated LASSO (L1 norm), ridge (L2 norm) and MCP [9] regularizations on Cox-PH models. CoxBoost is an iterative "gradient boosting" method modified from the Cox-PH model[10]. Finally, Random Forests Survival (RF-S) is a tree-based, non-linear, ensemble method [11], rather than a proportional hazards model.

To address all the issues of ANN based predictions as mentioned earlier, we have developed a new software package, named Cox-nnet. We used Cox regression as the output layer of ANN, rather than approximating it as a classification problem. We note three advantages of our package. First, it has statistically significant improved predictive accuracy in comparison with the other modeling methods described above. Secondly, it is also optimized for use with graphics processing units (GPU) and runs approximately 10 times faster compared to running on the central processing unit (CPU). Lastly, Cox-nnet allows feature importance scores to be defined, so that the relative importance of specific genes to prognosis outcome can be assessed. The hidden layer node structure in the ANN can be analyzed to reveal more useful information regarding relevant genes and pathways in the data. Overall, Cox-nnet is a desirable survival analysis method with both excellent predictive power and ability to elucidate biological functions related to prognosis.

5.4 Results

5.4.1 Cox-nnet structure and optimization

Cox-nnet is the neural network extension of the Cox-PH model. We created a package suitable for high dimensional datasets using the Theano Python library, a package created for the computation of mathematical expressions involving multi-dimensional arrays and machine learning [12]. The neural network model used in this paper is shown in Figure 1 and an overview of modules in the Cox-nnet package is shown in Figure S1. As a proof of concept, the current ANN architecture is composed of the input layer, one fully connected hidden layer and an output "proportional hazard" layer. The output layer of Cox-nnet replaces the linear predictors in the standard Cox-PH model (see theoretical descriptions in Methods).

Cox-nnet performs cross-validation (CV) to find the optimal regularization parameter. Due to the large amount of parameters and hyperparameters, overfitting is a potential problem in ANNs. Thus for regularization, we experimented with a range of regularization methods, including ridge, dropout [13], and the combination of ridge and dropout (see details in Methods). We found that dropout regularization offered overall the best model (Figure S2). Additionally, we compared Cox-nnet structures between one hidden layer and two hidden layers, and found that a single hidden layer Cox-nnet with only dropout regularization performed slightly better than that with two hidden layers (Figure S2). Thus, we used the single hidden-layer Cox-nnet with dropout regularization for comparison with other survival methods for any following analysis.

Many other functions are implemented to improve the usability of the package (Figure S1). Among then, the optimization strategies include momentum gradient descent [14] and Nesterov accelerated gradient [15]. A comparison of these descent methods is shown in Figure S3A, and we chose the best Nesterov accelerated gradient search method for this report. Other parameterization details of Cox-nnet are described in Methods. Moreover, this package can be run on multiple threads or a Graphics Processing Unit (GPU), and it achieves slightly faster training time compared to Random Forest and CoxBoost (Figure S3B). In all, Cox-nnet is a modern software implementation that can achieve efficient computational time.

5.4.2 Performance comparison of survival prediction methods

We compared four methods, including Cox-nnet, Cox-PH (including Ridge, LASSO and MCP penalizations), CoxBoost and RF-S on 10 datasets from The Cancer Genome Atlas (TCGA).

These datasets were selected for having at least 50 death events (Table S1). For each dataset, we trained the model on 80% of the samples, selected randomly, and determined the regularization parameter using 5-fold CV on the training set. We evaluated the performance on the remaining 20% holdout test set. We replicated this evaluation 10 times in order to assess the average performance of each method.

Three metrics are used to evaluate the performance of the model. The first one is Harrellas concordance index (C-index) calculated for censored survival data [16, 17]. It evaluates the relative ordering of the samples, comparing the prognostic index (i.e., log hazard ratio) of each patient with the survival times. The second metric is the inverse probability of censoring weighted (IPCW) estimate of the uncensored concordance [18]. This metric aims to overcome the inaccuracy of C-index when censoring time is correlated with the patientas hazard. The third metric is the log-ranked p-value from Kaplan-Meier survival curves of two different survival risk groups. This is done by using the median Prognosis Index (PI), the output of Cox-nnet, to dichotomize the patients into high risk and low risk groups, similar to our earlier reports [17, 19, 20]. A log-ranked p-value is then computed to differentiate the Kaplan-Meier survival curves from these two groups. Note, the dichotomization of patients ignores the differences within each dichotomized group, thus may lead to less accuracy compared to C-index and IPCW metrics.

The comparison of C-indices among the four methods over the 10 TCGA data are shown in Figure 2A. Overall, Cox-nnet (dropout) has higher predictive accuracy over the other three methods (Figure 2B). Cox-PH (ridge penalization) performs the second best, followed by CoxBoost and RF-S in descending order. Interestingly, the ensemble-based method RF-S consistently ranks worse than Cox-nnet and Cox-PH. The comparison of the IPCW metric and the log-ranked p-values on the dichotomized survival risk groups is shown in Figure S4 and S5 respectively. Generally, log-ranked p-values in the 10 TCGA datasets are better in Cox-nnet, compared to other methods. However, the dichotomization of patients ignores the differences within each dichotomized group, thus the resulting log-ranked p-values are less consistent than C-indices on the same data.

5.4.3 Hidden layer nodes of Cox-nnet are surrogate prognostic features

To explore the biological relevance of the hidden nodes of Cox-nnet, we used the TCGA Kidney Renal Cell Carcinoma (KIRC) dataset as an example. We first extracted the contribution of each hidden node to the PI score for each patient (Figure 3A). The contribution was calculated as the output value of each hidden node weighted by the corresponding coefficient at the Cox regression output layer. As expected, the value of the hidden nodes strongly correlated to the PI score.

However, there is still significant heterogeneity among the nodes, suggesting that individual nodes may reflect different biological processes. We hypothesize that the top (most variable) nodes may serve as surrogate features to discriminate patient survival. To explore this idea, we selected the top 20 nodes with the highest variances, and presented the patients PI scores using t-SNE (Figure 3B). t-SNE is a non-linear dimensionality reduction method that embeds high-dimensional datasets into a low dimensional space (usually two or three dimensions). This method has been widely used to visualize data with large number of features, by enhancing the separation among samples[21]. The hidden nodes represent a dimension reduction of the original data and they clearly discriminate samples by their PI scores, as shown by the t-SNE plot (Figure 3B, left). As comparison, we performed PCA using the whole RNA-Seq gene expression matrix, and then selected the top 20 PCA components with the highest degree of explained variance in the data. These 20 principle components from PCA in combination with t-SNE fail to separate the patient samples by PI score (Figure 3B, right). This drastic difference between the t-SNE plots demonstrates that the nodes in Cox-nnet effectively capture the survival information. Therefore, the top node PI scores can be used as features for dimension reduction in survival analysis.

5.4.4 Biological relevance of hidden layer nodes of Cox-nnet

To further explore the biological relevance of the top 20 hidden nodes, we conducted Gene Set Enrichment Analysis (GSEA) [22] using KEGG pathways [23], as described in the Methods section. Briefly, we calculated significantly enriched pathways using Pearsonas correlation between the log transformed gene expression to the output score of each node (Figure 3C and Table S2). We compared these enriched pathways to those from GSEA of the Cox-PH (ridge) model (Table S3), the competing model with the second best prognosis prediction. Using the fgsea package in R [24], we calculated statistical significance of the pathways by performing 10,000 permutations, followed by multiple hypothesis testing with Benjamini Hochberg adjustment. A total of 110 (out of 187) significantly enriched pathways (Table S2) were identified in at least one node, including seven pathways enriched in all 20 nodes that were not found by the Cox-PH method (Table 1). In contrast, Cox-PH only identified 30 significantly enriched pathways using the same significance threshold. We also used the genes values from CoxBoost or Random Forest, however they did not produce any significantly enriched pathways. Among the seven pathways enriched in all 20 nodes from Cox-nnet, the P53 signaling pathway stands out as an important biologically relevant pathway (Figure S6), since it was shown to be highly prognostic of patient survival in kidney cancer [25].

Next, we estimated the predicative accuracies of the leading edge genes [22] enriched in the GSEA from Cox-nnet vs. those enriched in Cox-PH model. Leading edge genes are those genes in the pathway of interest that contribute positively to the enrichment score in GSEA. We used the C-index of each leading edge gene, obtained from single-variable analysis (Figure 4). Collectively, leading edge genes from Cox-nnet have significantly higher C-index scores (p = 5.79e-05) than those from Cox-PH, suggesting that Cox-nnet has selected more informative features. In order to visualize these gene level and pathway level differences between Cox-nnet and Cox-PH, we reconstructed a bipartite graph between leading edge genes for Cox-nnet or feature genes (for Cox-PH) and their corresponding enriched pathways (Figure 5). Besides P53 pathway mentioned earlier that is specific to Cox-nnet, several other pathways, such as insulin signaling pathway, endocytosis and adherens junction, also have many more genes enriched in Cox-nnet. Among these genes specific to Cox-nnet, many have been previously reported to relevant to renal carcinoma development and prognosis, such as CASP9[26], TGFBR2[27], KDR (VEGFR)[28]. These results suggest that Cox-nnet model reveals richer biological information than Cox-PH.

5.4.5 Evaluation of gene input relative to survival in Cox-nnet

To further examine the importance of each gene relative to the survival outcome, we calculated the average partial derivative of each input gene feature value over all patients, with respect to the output of the model (i.e., the log hazard ratio). As demonstrated by the leading edge genes in seven common pathways of all nodes in Cox-nnet, the feature importance scores produce stronger biological insight (Figure S6). For example, the feature importance for the BAI1 gene in the P53 pathway is much higher in the Cox-nnet model compared to the Cox-PH model. Corresponding to our finding, the BAI gene family was found to be involved in several types of cancers including renal cancer[29] [30] [31] [32]. BAI1 acts as an inhibitor to angiogenesis and is transcriptionally regulated by P53 [33]. Its expression level was significantly decreased in tumor vs. normal kidney tissue, and was even lower in advanced stage renal carcinoma[32]. Mice kidney cancer models treated with BAI1 showed slower tumor growth and proliferation [34]. Additionally, the MAPK1 gene (also known as ERK2), annotated in two pathways identified by Cox-nnet (the Adherens Junction and Insulin Signaling pathway), has a much higher feature importance score in Cox-nnet compared to Cox-PH. MAPK1 is one of the key kinases in intracellular transduction, and was found constitutively activated in renal cell carcinoma [35]. Drugs inhibiting the MAPK cascade have been targeted for development[36].

Additionally, we compared the partial derivative of the hidden nodes (rather than the Cox-nnet output), with respect to the input genes (Supplemental Figure S7) and the partial derivative of the output with respect to the hidden nodes i.e., the output layer weights (Supplemental Figure S8). We first calculated the gradient for each patient and calculated the average partial derivatives, and replicated the GSEA analysis as for the previous analysis. However, we found that fewer pathways are significant, and are less relevant to cancer using this approach (Supplemental Figure S7). The authors of the GSEA algorithm stated that the input to the method should be the "correlation of the gene with the phenotype" [22] suggesting that the partial derivative approach may not be the appropriate metric.

5.5 Discussion

In this report, we have implemented Cox-nnet, a new ANN method, to predict patient survival from high throughput omics data. Cox-nnet is an improved alternative to the standard Cox-PH regression, as demonstrated by increased performance for survival prediction in 10 TCGA RNA-Seq datasets. Moreover, Cox-nnet has the capabilities to explore the biological information.

Our analysis suggests that Cox-nnet can reveal richer biological information than Cox-PH. This is manifested both at the pathway and gene levels. The hidden nodes in the Cox-nnet model have distinct activation patterns, and can serve as surrogate features for survival-sensitive dimension reduction. More significant KEGG pathways are enriched which correlate with top nodes in Cox-nnet, as compared to those from the Cox-PH model, suggesting that Cox-nnet reveals more relevant biological information. A critical pathway for renal cancer development, P53 pathway, is only enriched by Cox-nnet but not Cox-PH model in TCGA KIRC. Other pathways, including insulin signaling pathway, endocytosis and adherens junction, have many more genes enriched by Cox-nnet. Moreover, leading edge genes obtained from these KEGG pathways enriched by Cox-nnet (which are a fraction of the gene features considered by the model) have collectively higher associations with survival.

Some technical details on model optimization is worth discussion. In neural networks, because of the large amount of parameters and hyperparameters, overfitting is a potential problem. In Cox-nnet, we experimented with three regularization approaches given previous guidelines: ridge, dropout and combination of ridge and dropout. Ridge regularization is one of the most common methods to reduce overfitting, recommended by Demuth et al. [37]. In this scheme, the L2 norm of all the weights are added to the cost function of the model, leading to a "weight decay" term in the gradient. Dropout is a recent regularization method for networks, inspired

by Bayesian analysis on weighted averages of different network architectures to improve the model performance [13]. In dropout networks, each training iteration uses a different network architecture; nodes are randomly removed from the network during training based on a probability hyperparameter between 0 and 1. Instead of entire models being reweighted, the output of each node is reweighted during evaluation. This method was previously shown to perform better than other regularization methods, such as ridge regularization [13]. Our results on Cox-nnet confirmed this earlier conclusion. Also similar to previous study, we found that additional complexity of combining dropout and ridge regularization does not yield better performance [13].

Another potential risk of overfitting in Cox-nnet can come from inadvertently fitting the model to the patient censoring. To investigate this, we ran a simulation RNA-Seq dataset (described in the methods section) and compared C-index and IPCW metrics with censoring to uncensored C-index (Figure S9). The uncensored performance index represents the "true" performance of the model, i.e., if all the data were uncensored. Both the C-index and the IPCW metric accurately estimate the uncensored performance of the simulated dataset, and neural network-based Cox-nnet and tree-based Random Forest survival do not differ significantly from Cox-PH models. These results suggest that overfitting may not be a significant concern in Cox-nnet, which could have inadvertently benefited from overtuning of the hyperparameters.

As a promising new predictive method for prognosis, the current Cox-nnet implementation has some limitations. Its architecture is relatively simple, including an input layer, one or two hidden layer and an output Cox regression layer. It is possible to incorporate other more sophisticated architecture into the model, such as including more layers of neurons or more sophisticated hidden layers (although the size of current genomics data suggest deeper ANN is not necessarily more beneficial, Figure S2). A convolutional neural network approach using convolutional and pooling layers could also be used, as those reported in processing imaging or other types of positional data [38]. Additionally, it is possible to embed a priori biological pathway information into the network architecture, e.g., by connecting genes in a pathway to a common node in the next hidden layer of neurons. In the future, we plan to further analyze how different neural network architectures affect the performance of Cox-nnet and compare the biological insights from the various models.

5.6 Methods

5.6.1 Cox-PH, CoxBoost and Random Forest Survival (RF-S) models

Cox-nnet is an extension to the Cox-PH model. Individual hazard, an instantaneous measure of the likelihood of an event, is estimated based on a set of features [8]. The hazard function is:

$$h(t|x_i) = h_0(t)exp(\theta_i) \tag{12}$$

$$\theta_i = x_i^T \beta \tag{13}$$

Where θ_i is the log hazard ratio for patient i. This model uses partial log-likelihood as the cost function:

$$PL(\beta) = \prod_{C(i)=1} \frac{exp\theta_i}{\sum_{t_j \geq t_i} exp\theta_j} \tag{14}$$

CoxBoost, is an iterative "gradient boosting" method modified from the Cox-PH model [10]. In CoxBoost, parameters are separated into individual partitions, and the partition that leads to the largest improvement in the penalized partial likelihood is selected for that iteration. In subsequent boosting iteration, the model selects another block and refits those parameters by maximizing the penalized likelihood function. In this method, the number of boosting iterations is used as the complexity parameter in CoxBoost and optimized via cross-validation (CV).

Random Forests Survival (RF-S) is a tree-based, non-linear, ensemble method [11], rather than a proportional hazards model. For each tree in the forest, data are bootstrapped, and nodes are split by maximizing the log-rank statistic. At the leaf nodes, the cumulative hazard function (CHF) is estimated and a patientas CHF is calculated as an average over all the trees in the ensemble.

5.6.2 Theoretical considerations of Cox-nnet

Cox-nnet is a neural network whose output layer is replaced by a Cox model. In a Cox-nnet model with one input layer of J input features and one hidden layer composed of H hidden nodes, the linear predictor is replaced by the outputs of the hidden layer:

$$\theta_i = G(W^T X_i + b)^T \beta \tag{15}$$

Where W is the coefficient weight matrix between the input and hidden layer with the size H x J, is the bias term for each hidden node and is the activation function (applied element-wise on a vector). Subsequently, the ridge (L2 norm) regression cost function is modified to:

$$Cost(\beta) = -pl(\beta) + \lambda \|\beta\|^2 \tag{16}$$

In this manuscript, the tanh activation function is used:

$$G(z) = \frac{exp(z) - exp(-z)}{exp(z) + exp(-z)} \tag{17}$$

In addition to ridge regularization, we also employ dropout regularization [13]. This approach has been shown to reduce overfitting and improve performance over other regularization schemes[13]. In dropout, nodes are removed during each training iteration with probability 1-p. During evaluation, output from the nodes are multiplied by p. The optimal dropout parameter, p, is determined through CV on the training set. Moreover, we also apply the combined approach of ridge and dropout for comparison.

5.6.3 Implementation of Cox-nnet

We implement Cox-nnet using a feed forward, back propagation network with gradient descent. The partial log likelihood is usually written as a double conditional sum (equation 3). To avoid the computational inefficiency of calculating the partial log likelihood (equation 3) using two nested for loops, we convert it into a formulation of matrix operations and basic sums. First we define an indicator matrix R with elements:

$$R_{ij} = \begin{cases} 1 \ if \ t_i \leq t_j \\ 0 \ if \ t_i > t_j \end{cases} \tag{18}$$

We also define an indicator vector C with elements given by the censoring of each patient. An operation using R replaces the conditional sum over $t_j \geq t_i$, and an operation using C replaces the conditional sum over C(i)=1 in equation 3. In Theano, the partial log likelihood is:

$$pl = T.sum((theta - T.log(T.sum(T.exp(theta) * R, axis = 1))) * C) \tag{19}$$

For the models trained in this manuscript, the number of iterations was fixed at 1e4. The learning rate was initialized at 0.1, and decayed exponentially by a factor of 0.9 if the loss did not decrease. The number of hidden nodes in the hidden layer is chosen to be the square root of the number of input nodes, following the aoepyramida rule of thumb [39]. The optimization strategies used was Nesterov accelerated gradient [15].

Many functions are implemented to improve the usability of the package, including CVSearch, CVProfile, CrossValidation, and TrainCoxMlp (Figure S1). CVSearch, CVProfile, CrossValidation are methods that perform CV to find the optimal regularization parameter. TrainCoxMlp performs optimization of coefficients on the regularized partial likelihood function.

The source code of cox-nnet can be found at: `https://github.com/lanagarmire/cox-nnet`, and can be installed through the Python Package Index (PyPI). Documentation of package can be found at `http://garmiregroup.org/cox-nnet/docs/`. To call Cox-nnet from R, we provide an example here: `http://garmiregroup.org/cox-nnet/docs/examples/#interfacing-and-analysis-with-r`.

5.6.4 Model evaluation

To evaluate the performance of all methods, we resampled the data 10 times. In each resampling iteration, we trained each model on 80% of the samples for each dataset (chosen randomly) and evaluated the performance on the 20% holdout test set. The output of Cox-PH, Cox-nnet and CoxBoost are the log hazard ratios (i.e., Prognosis Index, or PI) for each patient. The hazard ratio describes the relative risk of a patient compared to a non-parametric baseline. On the other hand, the output of RF-S is an estimation of the survival time for each patient. We use C-index, IPCW [18] and log-ranked p-values to measure the performance of each model.

C-index: is a measure of how well the model prediction corresponds to the ranking of the survival data [40]. It is calculated for censored survival data, which evaluates a value between 0 and 1, with 0.5 equivalent to a random process. The C-index can be computed as a summation over all events in the dataset, where patients with a higher survival time and lower log hazard ratios (and conversely patients with a lower survival time but higher log hazard ratios) are considered concordant. IPCW: This metric aims to overcome the inaccuracy of C-index when censoring time is correlated with the patientas hazard.

log-ranked p-value: a PI cutoff threshold is used to dichotomize the patients in the data set into higher and lower risk groups, similar to our earlier report [19, 20]. A log-ranked p-value is then computed to differentiate the Kaplan-Meier survival curves between the higher vs. lower risk groups. In this report, we used the median log hazard ratio as the cutoff threshold.

The cross-validated performance metric may be Harrel's concordance index (C-index) [16] or the "cross-validated partial likelihood" [41]. Since the contribution of each patient in the partial likelihood is determined only in the context of all the other patients, the cross-validated partial likelihood is calculated subtracting full partial likelihood from the training set in the CV. In the k-th iteration of a K-fold CV, the optimal coefficients $\hat{\beta}_{\lambda,k}$ are found by minimizing the cost function on the training sub-samples. If $pl_k(\hat{\beta}_{\lambda,k})$ is the partial likelihood of the training sub-samples, and $pl(\hat{\beta}_{\lambda,k})$ is the partial likelihood of the full dataset, then the cross-validated partial likelihood is the sum of differences:

$$cvpl(\lambda) = \sum_{k=1}^{K} pl(\hat{\beta}_{\lambda,k}) - pl_k(\hat{\beta}_{\lambda,k}) \tag{20}$$

5.6.5 Feature evaluation

For computing the importance of a feature in Cox-nnet, we use a method of partial derivatives [42, 43]. For each patient, we compute the partial derivatives of each input with respect to the linear output of the model (e.g., the log hazard ratio). The average of the partial derivatives for each input across all patient samples is calculated as the feature score.

5.6.6 Datasets

In order to evaluate the performance of Cox-nnet, we analyzed 10 TCGA datasets which were combined into a pan-cancer dataset. The TCGA datasets included the following cancer types: Bladder Urothelial Carcinoma (BLCA), Breast invasive carcinoma (BRCA), Head and Neck squamous cell carcinoma (HNSC), Kidney renal clear cell carcinoma (KIRC), Brain Lower Grade Glioma (LGG), Liver hepatocellular carcinoma (LIHC), Lung adenocarcinoma (LUAD), Lung squamous cell carcinoma (LUSC), Ovarian serous cystadenocarcinoma (OV) and Stomach adenocarcinoma (STAD). RNA-Seq expression and clinical data were downloaded from the Broad Institute GDAC [44]. Overall survival time and censoring information were extracted from the clinical follow-up data. Raw count data were normalized using the DESeq2 R package [45] and then log-transformed. Datasets were selected from TCGA based on the following

criteria: $>$ 300 samples with both RNASeq and survival data and $>$ 50 survival events. In total, 5031 patient samples were used (see Table S1 for a patient tabulation by individual dataset).

5.6.7 t-SNE clustering

T-distributed stochastic neighbor embedding (t-SNE) is a non-linear dimensionality reduction method that embeds high-dimensional datasets into a low dimensional (usually two or three) space. Comparing to other methods with similar purposes (such as MDS), t-SNE puts more emphasis on making sure data points that are close together in the original high-dimensional space remain close in the dimension-reduced space. Therefore it usually preserves the clusters of data points in the original space [21]. This method has been widely used to visualize data with large number of features. To explore this idea, we selected the top 20 nodes of the Cox-nnet model with the highest variances, and clustered the patient samples using t-SNE. To do this, we used the tsne package in R [46].

5.6.8 Statistical testing between model performance

To test for statistical significance between methods using their performance metrics (C-index, IPCW and log-rank), we use a one-tailed Wilcoxon rank sum test for each of the 10 TCGA datasets, to compare Cox-nnet with the other methods. The p-value is adjusted using the Benjamini Hochberg procedure.

5.6.9 Data simulation

We used the ssizeRNA package in R to generate simulated RNA-Seq data counts in R [47]. We generated four sub-groups of 200 patients each (a total of 800 patient samples) with 1000 genes, with 20% of the genes differentially expressed for each group. The prognosis index for patients in each group were randomly generated based on expression of 100 randomly selected genes, and the survival times were sampled based on the Weibull survival distribution. Censoring times were chosen from the exponential distribution with rate = 0.05. We randomly generated this dataset 100 times and estimated the performance metrics on 20% holdout test-sets.

5.7 Acknowledgements

This research was supported by grants K01ES025434 awarded by NIEHS through funds provided by the trans-NIH Big Data to Knowledge (BD2K) initiative (www.bd2k.nih.gov), P20 COBRE GM103457 awarded by NIH/NIGMS, R01 LM012373 awarded by NLM, R01 HD084633 awarded by NICHD, and Hawaii Community Foundation Medical Research Grant 14ADVC-64566 to L.X. Garmire.

References

1. Parsell, M. Steven M. Platek, Julian Paul Keenan and Todd K. Shackelford (eds), Evolutionary Cognitive Neuroscience. *Minds and Machines* **19,** 275–278. ISSN: 0924-6495 (2009).
2. McCulloch, W. S. & Pitts, W. A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics* **5,** 115–133. ISSN: 0007-4985 (1943).
3. Jones, N. The learning machines. *Nature* (2014).
4. Faraggi, D. & Simon, R. A neural network model for survival data. *Statistics in medicine* **14,** 73–82. ISSN: 1097-0258 (1995).
5. Chi, C.-L., Street, W. N. & Wolberg, W. H. Application of artificial neural network-based survival analysis on two breast cancer datasets. *AMIA Symposium* **2007,** 130 (2007).
6. Petalidis, L. P., Oulas, A., Backlund, M., Wayland, M. T., Liu, L., Plant, K., Happerfield, L., Freeman, T. C., Poirazi, P. & Collins, V. P. Improved grading and survival prediction of human astrocytic brain tumors by artificial neural network analysis of gene expression microarray data. *Molecular cancer therapeutics* **7,** 1013–1024. ISSN: 1535-7163 (2008).
7. Joshi, R. & Reeves, C. *Beyond the Cox model: artificial neural networks for survival analysis part II* in *Proceedings of the eighteenth international conference on systems engineering* (2006), 179–184.
8. Therneau, T. M. & Grambsch, P. M. *Modeling survival data: extending the Cox model* ISBN: 1475732945 (Springer Science & Business Media, 2000).
9. Breheny, P. & Huang, J. Coordinate descent algorithms for nonconvex penalized regression, with applications to biological feature selection. *The annals of applied statistics* **5,** 232 (2011).
10. Binder, H. CoxBoost: Cox models by likelihood based boosting for a single survival endpoint or competing risks. *R package version* **1** (2013).

11. Ishwaran, H., Kogalur, U. B., Blackstone, E. H. & Lauer, M. S. Random survival forests. *The Annals of Applied Statistics,* 841–860. ISSN: 1932-6157 (2008).

12. Al-Rfou, R., Alain, G., Almahairi, A., Angermueller, C., Bahdanau, D., Ballas, N., Bastien, F., Bayer, J., Belikov, A. & Belopolsky, A. Theano: A Python framework for fast computation of mathematical expressions. *arXiv preprint arXiv:1605.02688* (2016).

13. Srivastava, N., Hinton, G. E., Krizhevsky, A., Sutskever, I. & Salakhutdinov, R. Dropout: a simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research* **15,** 1929–1958 (2014).

14. Qian, N. On the momentum term in gradient descent learning algorithms. *Neural networks* **12,** 145–151. ISSN: 0893-6080 (1999).

15. Bengio, Y., Boulanger-Lewandowski, N. & Pascanu, R. *Advances in optimizing recurrent networks* in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on* (IEEE, 2013), 8624–8628. ISBN: 1520-6149.

16. Koziol, J. A. & Jia, Z. The concordance index C and the Mann-Whitney parameter Pr (X¿ Y) with randomly censored data. *Biometrical Journal* **51,** 467–474. ISSN: 1521-4036 (2009).

17. Wei, R., De Vivo, I., Huang, S., Zhu, X., Risch, H., Moore, J. H., Yu, H. & Garmire, L. X. Meta-dimensional data integration identifies critical pathways for susceptibility, tumorigenesis and progression of endometrial cancer. *Oncotarget.* ISSN: 1949-2553 (2016).

18. Gerds, T. A., Kattan, M. W., Schumacher, M. & Yu, C. Estimating a time-dependent concordance index for survival prediction models with covariate dependent censoring. *Statistics in Medicine* **32,** 2173–2184. ISSN: 1097-0258 (2013).

19. Huang, S., Chong, N., Lewis, N. E., Jia, W., Xie, G. & Garmire, L. X. Novel personalized pathway-based metabolomics models reveal key metabolic pathways for breast cancer diagnosis. *Genome medicine* **8,** 1. ISSN: 1756-994X (2016).

20. Huang, S., Yee, C., Ching, T., Yu, H. & Garmire, L. X. A Novel Model to Combine Clinical and Pathway-Based Transcriptomic Information for the Prognosis Prediction of Breast Cancer. *PLoS computational biology* **10,** e1003851. ISSN: 1553-7358 (2014).

21. Maaten, L. v. d. & Hinton, G. Visualizing data using t-SNE. *Journal of Machine Learning Research* **9,** 2579–2605 (2008).

22. Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., Paulovich, A., Pomeroy, S. L., Golub, T. R. & Lander, E. S. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences* **102,** 15545–15550. ISSN: 0027-8424 (2005).

23. Kanehisa, M. & Goto, S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic acids research* **28,** 27–30. ISSN: 0305-1048 (2000).

24. Sergushichev, A. An algorithm for fast preranked gene set enrichment analysis using cumulative statistic calculation. *bioRxiv,* 060012 (2016).

25. Girgin, C., Tarhan, H., seyin uuml, u., Hekimgil, M., Sezer, A. & Gurel, G. P53 mutations and other prognostic factors of renal cell carcinoma. *Urologia internationalis* **66,** 78–83. ISSN: 1423-0399 (2001).

26. Marques, I., Teixeira, A. L., Ferreira, M., Assis, J., Lobo, F., Mauricio, J. & Medeiros, R. Influence of survivin (BIRC5) and caspase-9 (CASP9) functional polymorphisms in renal cell carcinoma development: a study in a southern European population. *Molecular biology reports* **40,** 4819–4826. ISSN: 0301-4851 (2013).

27. Akhurst, R. J. & Derynck, R. TGF-Beta signaling in cancer-a double-edged sword. *Trends in cell biology* **11,** S44–S51. ISSN: 0962-8924 (2001).

28. Choueiri, T. K., Vaishampayan, U., Rosenberg, J. E., Logan, T. F., Harzstark, A. L., Bukowski, R. M., Rini, B. I., Srinivas, S., Stein, M. N. & Adams, L. M. Phase II and biomarker study of the dual MET/VEGFR2 inhibitor foretinib in patients with papillary renal cell carcinoma. *Journal of Clinical Oncology* **31,** 181–186. ISSN: 0732-183X (2013).

29. Cork, S. M. & Van Meir, E. G. Emerging roles for the BAI1 protein family in the regulation of phagocytosis, synaptogenesis, neurovasculature, and tumor development. *Journal of molecular medicine* **89,** 743–752. ISSN: 0946-2716 (2011).

30. Fukushima, Y., Oshika, Y., Tsuchida, T., Tokunaga, T., Hatanaka, H., Kijima, H., Yamazaki, H., Ueyama, Y., Tamaoki, N. & Nakamura, M. Brain-specific angiogenesis inhibitor 1 expression is inversely correlated with vascularity and distant metastasis of colorectal cancer. *International journal of oncology* **13,** 967–970. ISSN: 1019-6439 (1998).

31. Lee, J., Koh, J., Shin, B., Ahn, K., Roh, J., Kim, Y. & Kim, K. K. Comparative study of angiostatic and anti-invasive gene expressions as prognostic factors in gastric cancer. *International journal of oncology* **18,** 355–362. ISSN: 1019-6439 (2001).

32. Izutsu, T., Konda, R., Sugimura, J., Iwasaki, K. & Fujioka, T. Brain-specific angiogenesis inhibitor 1 is a putative factor for inhibition of neovascular formation in renal cell carcinoma. *The Journal of urology* **185,** 2353–2358. ISSN: 0022-5347 (2011).

33. Nishimori, H., Shiratsuchi, T., Urano, T., Kimura, Y., Kiyono, K., Tatsumi, K., Yoshida, S., Ono, M., Kuwano, M. & Nakamura, Y. A novel brain-specific p53-target gene, BAI1, containing thrombospondin type 1 repeats inhibits experimental angiogenesis. *Oncogene* **15,** 2145–2150. ISSN: 0950-9232 (1997).

34. Kudo, S., Konda, R., Obara, W., Kudo, D., Tani, K., Nakamura, Y. & Fujioka, T. Inhibition of tumor growth through suppression of angiogenesis by brain-specific angiogenesis inhibitor 1 gene transfer in murine renal cell carcinoma. *Oncology reports* **18,** 785–792. ISSN: 1021-335X (2007).

35. Oka, H., Chatani, Y., Hoshino, R., Ogawa, O., Kakehi, Y., Terachi, T., Okada, Y., Kawaichi, M., Kohno, M. & Yoshida, O. Constitutive activation of mitogen-activated protein (MAP) kinases in human renal cell carcinoma. *Cancer research* **55,** 4182–4187. ISSN: 0008-5472 (1995).

36. Friday, B. B. & Adjei, A. A. Advances in targeting the Ras/Raf/MEK/Erk mitogen-activated protein kinase cascade with MEK inhibitors for cancer therapy. *Clinical Cancer Research* **14,** 342–346. ISSN: 1078-0432 (2008).

37. Demuth, H. B., Beale, M. H., De Jess, O. & Hagan, M. T. *Neural network design* ISBN: 0971732116 (Martin Hagan, 2014).

38. LeCun, Y. & Bengio, Y. Convolutional networks for images, speech, and time series. *The handbook of brain theory and neural networks* **3361,** 1995 (1995).

39. Masters, T. *Practical neural network recipes in C++* ISBN: 0124790402 (Morgan Kaufmann, 1993).

40. Harrell, F. E., Lee, K. L. & Mark, D. B. Tutorial in biostatistics multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Statistics in medicine* **15,** 361–387. ISSN: 0277-6715 (1996).

41. Van Houwelingen, H. C., Bruinsma, T., Hart, A. A., van't Veer, L. J. & Wessels, L. F. Cross-validated Cox regression on microarray gene expression data. *Statistics in medicine* **25,** 3201–3216. ISSN: 1097-0258 (2006).

42. Gevrey, M., Dimopoulos, I. & Lek, S. Review and comparison of methods to study the contribution of variables in artificial neural network models. *Ecological modelling* **160,** 249–264. ISSN: 0304-3800 (2003).

43. Olden, J. D., Joy, M. K. & Death, R. G. An accurate comparison of methods for quantifying variable importance in artificial neural networks using simulated data. *Ecological Modelling* **178,** 389–397. ISSN: 0304-3800 (2004).

44. Broad. Broad Institute TCGA Genome Data Analysis Center (2014): Analysis Overview for 15 July 2014. *Broad Institute of MIT and Harvard.* doi:10.7908/C1DN43V9 (2014).

45. Love, M., Anders, S. & Huber, W. Differential analysis of RNA-Seq data at the gene level using the DESeq2 package. *Bioconductor* (2013).

46. Donaldson, J. & Donaldson, M. J. Package 'tsne'. *CRAN Repository* (2010).

47. Bi, R., Liu, P. & Triche, T. Package 'ssizeRNA'. *Bioconductpor* (2016).

Table 1. Cox-nnet node-associated pathways. Significantly enriched pathways from common to all 20 hidden nodes that are not found in the Cox-PH Gene Set Enrichment Analysis (Adjusted p <0.05).

Pathway	P.value	P.adjusted	Nodes
KEGG adherens junction	0.000	0.001	1A-20
KEGG endocytosis	0.000	0.001	1A-20
KEGG insulin signaling pathway	0.000	0.001	1A-20
KEGG lysine degradation	0.000	0.003	1A-20
KEGG p53 signaling pathway	0.000	0.003	1A-20
KEGG pyruvate metabolism	0.000	0.001	1A-20
KEGG sphingolipid metabolism	0.001	0.005	1A-20

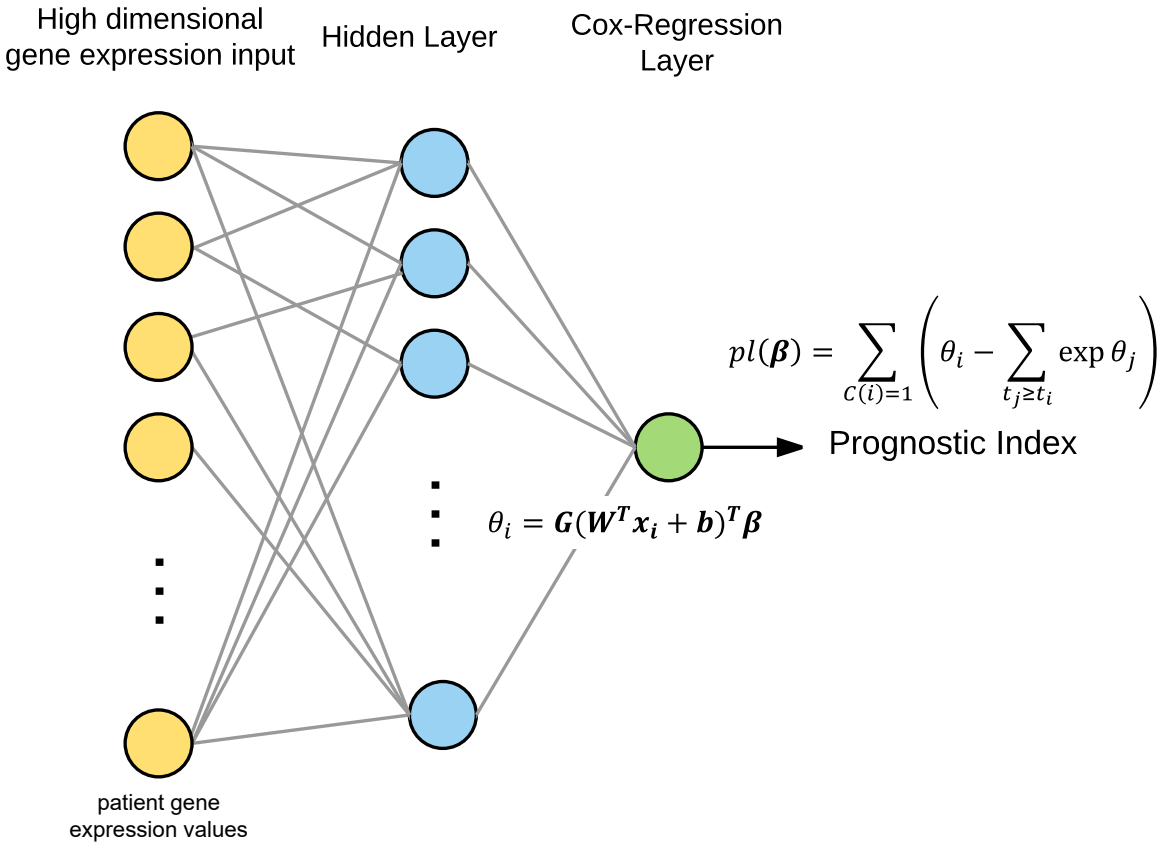


Figure 1. An overview of the neural network architecture used in this study.

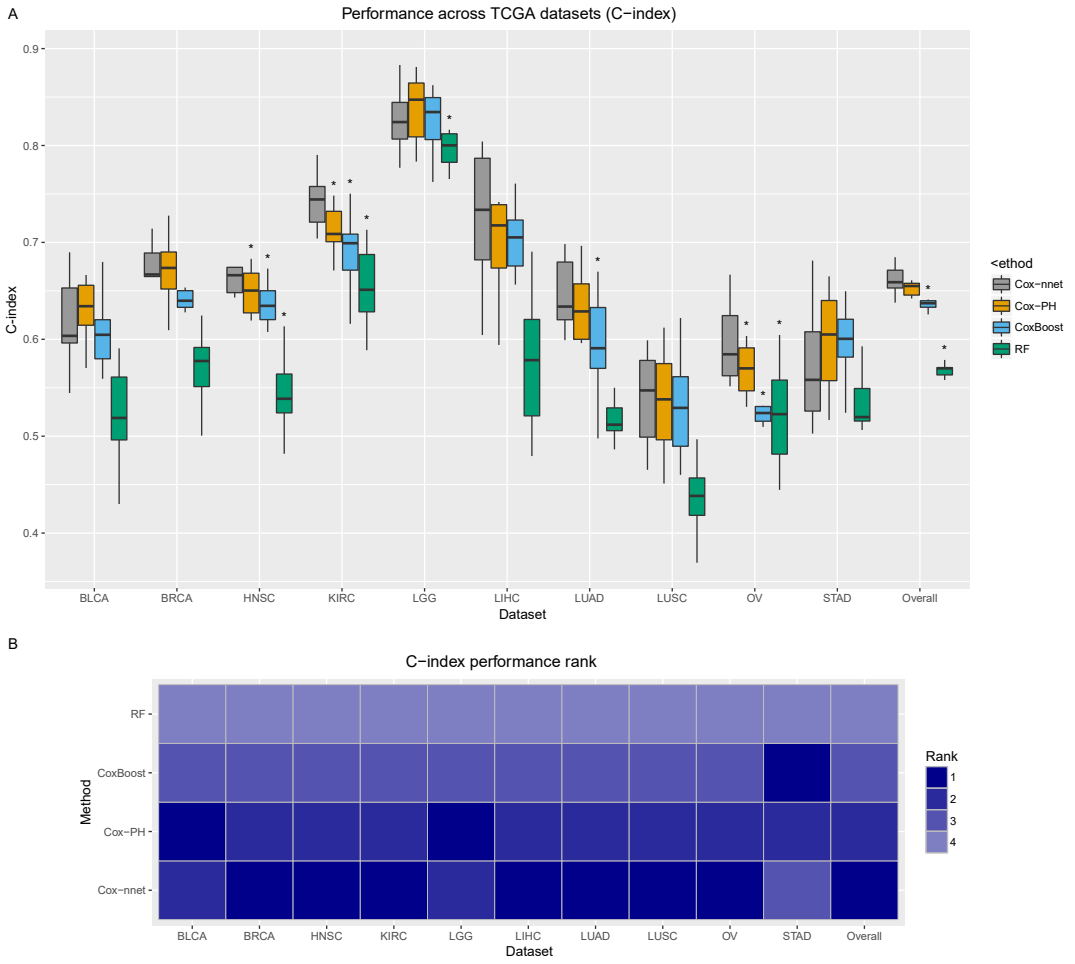


Figure 2. A. Boxplot of the C-index of the 10 TCGA datasets using four prognosis-predicting methods (Cox-nnet, CoxBoost, Cox-PH and RF-S). Each dataset was randomly split into 80% training and 20% testing sets. B. Heatmap of the performance rank of each dataset.

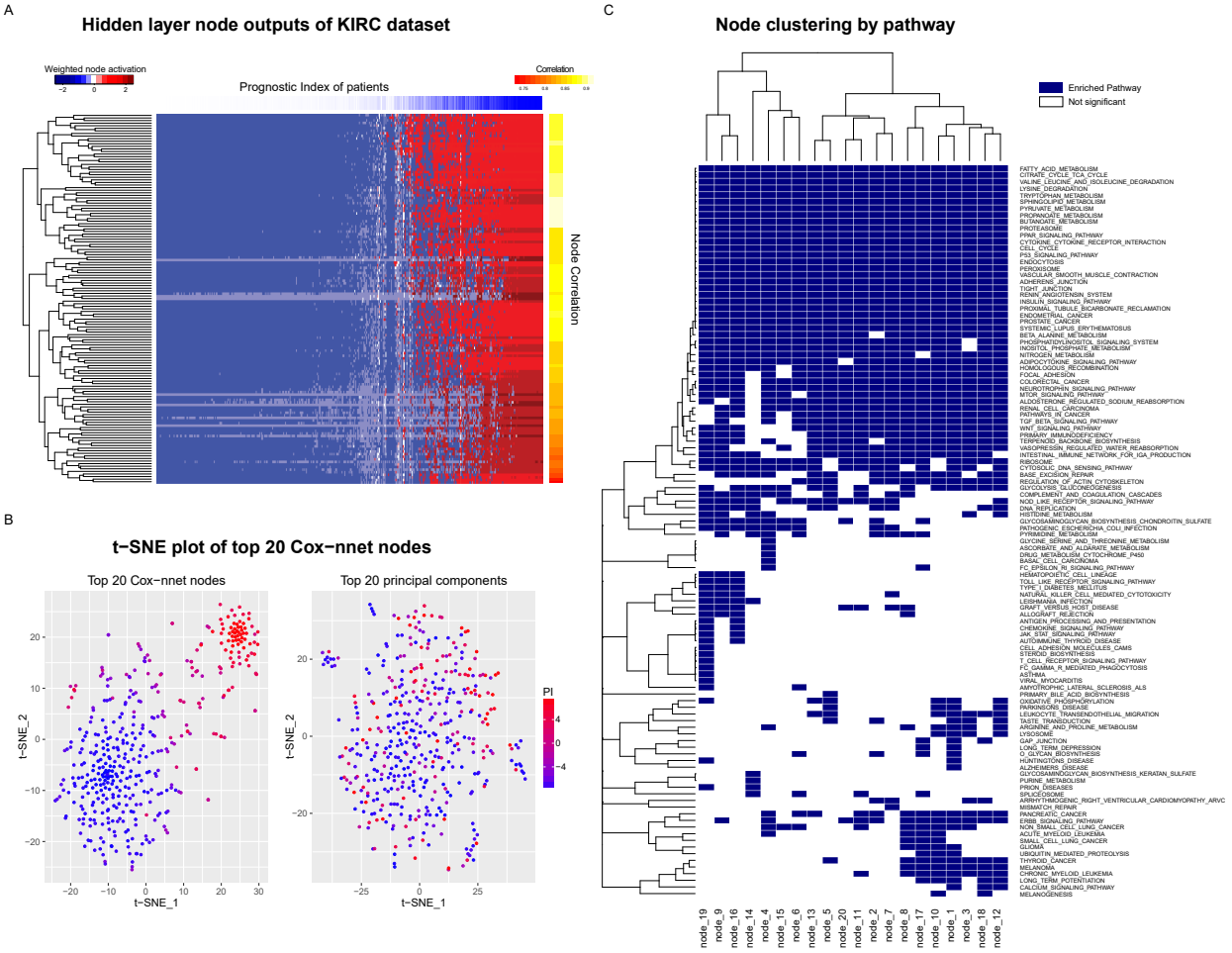


Figure 3. A. Hidden node activation weighted by the corresponding Cox layer coefficients of the TCGA KIRC dataset. The columns represent individual patient scores, ordered by their Prognostic Index. The rows represent the node activations. B. t-SNE plot of the top 20 hidden nodes and the top 20 principal components in PCA analysis of the RNA-Seq gene expressiongse. C. Gene Set Enrichment Analysis: significantly enriched KEGG pathways of the top 20 hidden nodes (adjusted p-value <0.05).

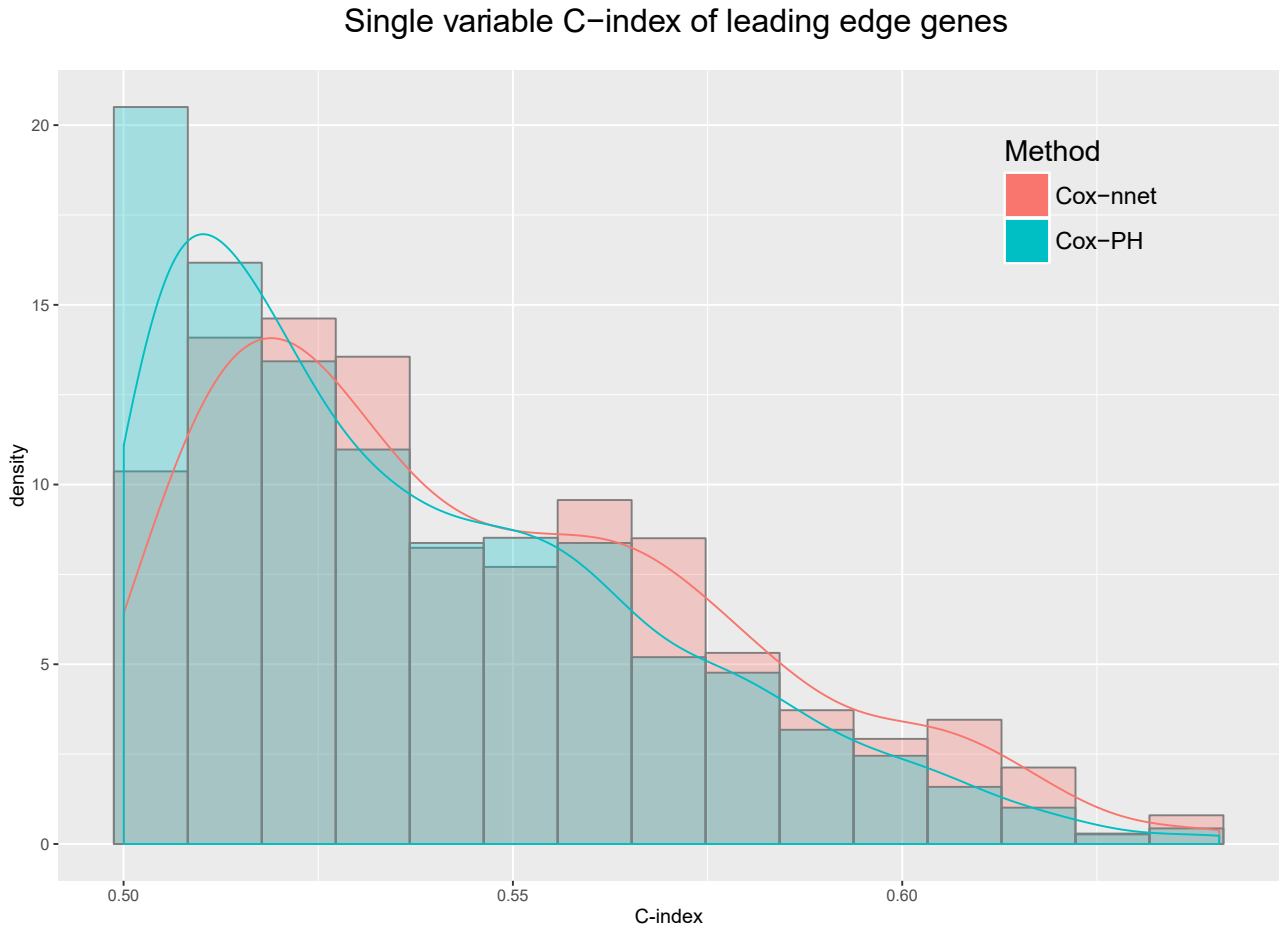


Figure 4. Single variable C-index scores of the leading edge genes from Cox-nnet and Cox-PH. Cox-nnet has significantly higher C-index scores (p = 5.79e-5).

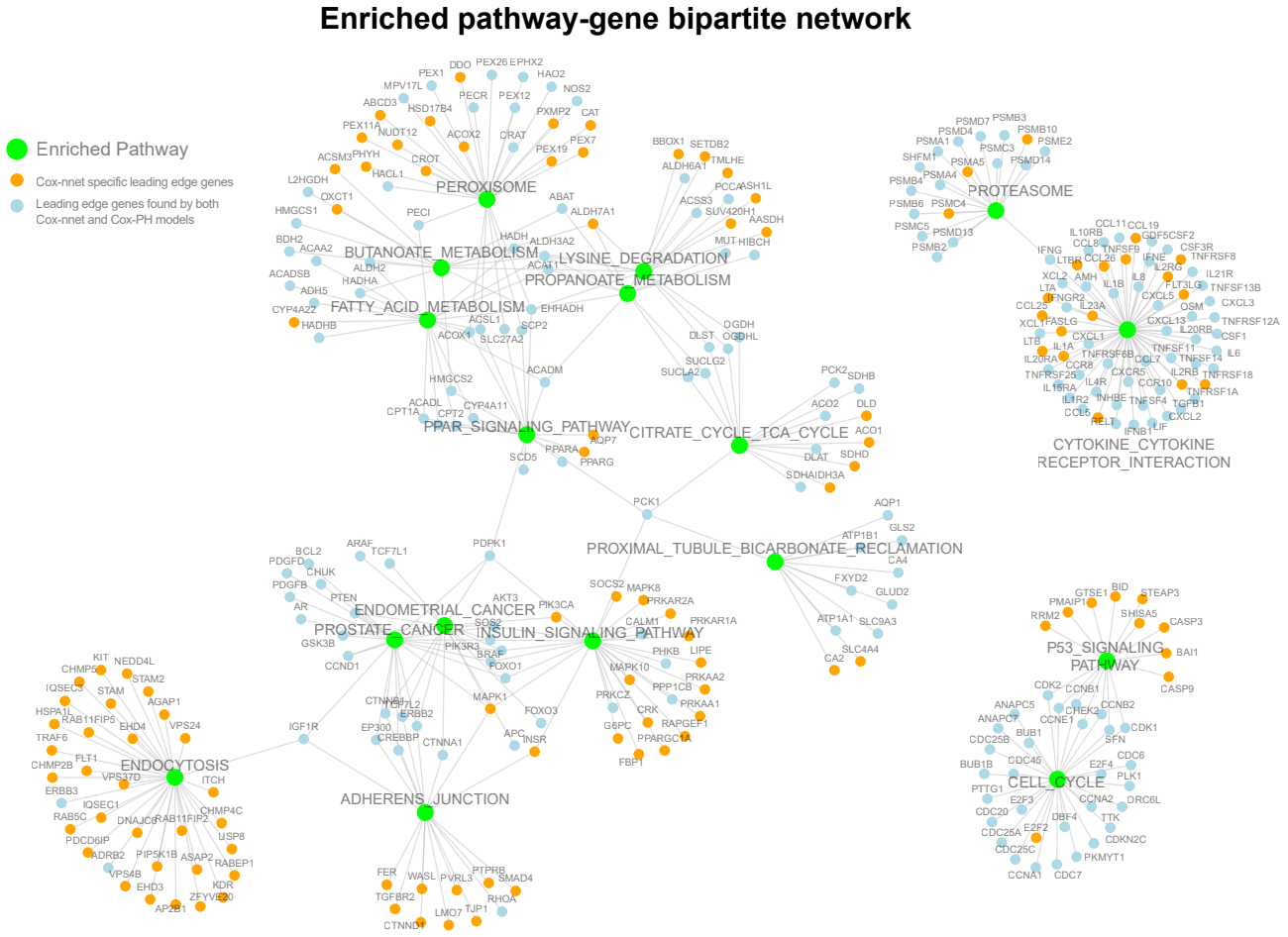


Figure 5. Enriched pathway-gene bipartite network from the leading edge genes and significantly enriched pathways. Significantly enriched pathways common to all 20 hidden nodes are labeled in green. Leading edge genes found uniquely in Cox-nnet are labeled in orange, and genes found in both Cox-nnet and Cox-PH are labeled in blue.

5.8 Appendix

5.8.1 Supplemental figures and tables

Figure S1. Overview of the structure, methods and classes in the Cox-nnet package.

Figure S2. A. Boxplot of the C-index of the 10 TCGA datasets among various penalization approaches in Cox-nnet one hidden layer (ridge, drop-out and ridge combined with dropout), Cox-PH (LASSO, ridge, and MCP), as well as Cox-nnet with two hidden layers. Each dataset was randomly split into 80% training and 20% testing sets and resampled 10 times to calculate the average performance of each approach.

Figure S3. A: comparison of descent methods on the TCGA KIRC dataset. The change in cost function is evaluated over 100,000 iterations for three methods: gradient descent, momentum gradient descent and the Nesterov accelerated gradient. B: Training time comparing CPU training time vs. GPU training time on the same dataset.

Figure S4. inverse probability of censoring weighted (IPCW) performance metric boxplot of the 10 TCGA datasets.

Figure S5. A. Bar plots of Log-rank p-values of the 10 TCGA datasets. The log rank p-values were calculated first splitting the patients by median prognostic index in the testing data set, and subsequently log-rank tests were performed to compare the survival distributions between the high and low risk groups. B. Kaplan-Meier plots showing survival differences between the high and low risk groups.

Figure S6. Variable importance of the common leading edge genes of enriched KEGG pathways.

Figure S7. Pathway enrichment using partial derivatives of the hidden nodes. Rather than using gene input correlation to the hidden node output, the partial derivative of the hidden nodes with respect to the input gene were calculated. Geneset Enrichment Aanalysis was used in the same manner as in Figure 3B. Using this approach, fewer significant pathways were detected.

Figure S8. Partial derivative of the output with respect to the hidden nodes compared to the hidden node output.

Figure S9. RNA-Seq survival simulation results showing the performance over 100 simulated datasets comparing the C-index, IPCW metric and the uncensored concordance.

Table S1. Tabulation of TCGA patients by individual dataset.

Table S2. Significantly enriched pathways from the Cox-PH method ($p < 0.05$).

Table S3. Significantly enriched pathways from the Cox-nnet method ($p < 0.05$).

Figure S1

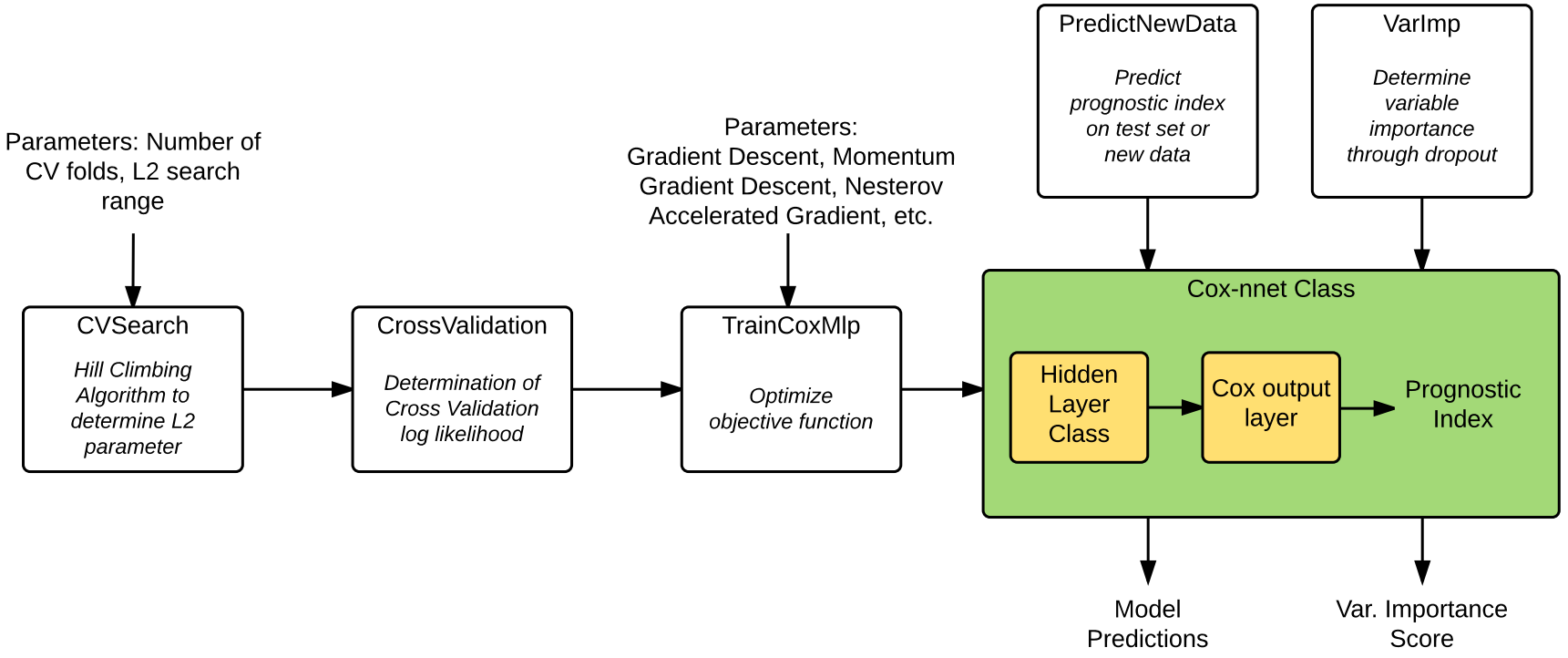

Parameters: Number of CV folds, L2 search range
CVSearch
Hill Climbing Algorithm to determine L2 parameter
CrossValidation
Determination of Cross Validation log likelihood
Parameters: Gradient Descent, Momentum Gradient Descent, Nesterov Accelerated Gradient, etc.
TrainCoxMlp
Optimize objective function
PredictNewData
Predict prognostic index on test set or new data
VarImp
Determine variable importance through dropout
Cox-nnet Class
Hidden Layer Class
Cox output layer
Prognostic Index
Model Predictions
Var. Importance Score

Figure S2

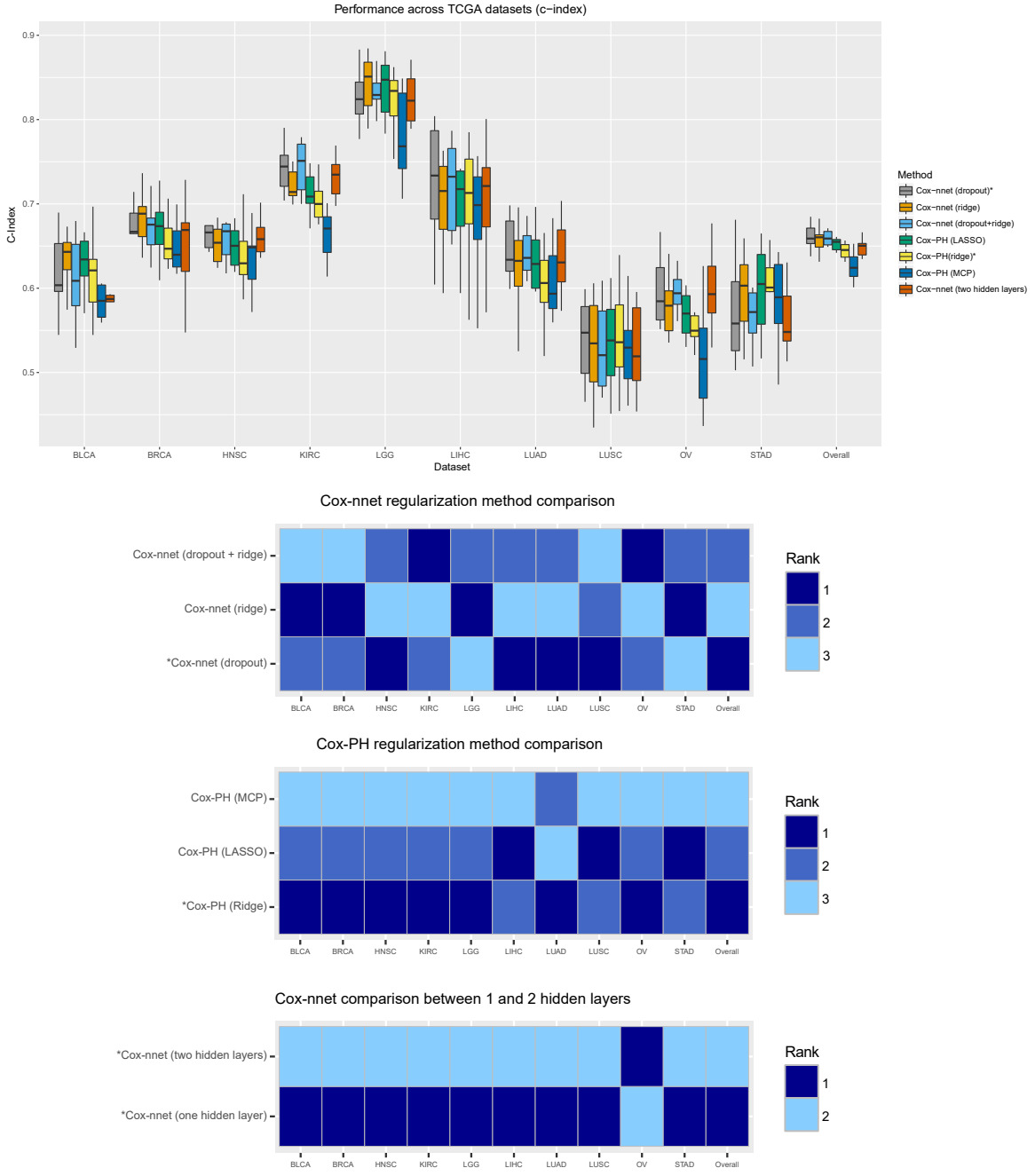

Performance across TCGA datasets (c−index)
C-Index
0.9
0.8
0.7
0.6
0.5
BLCA
BRCA
HNSC
KIRC
LGG
LIHC
LUAD
LUSC
OV
STAD
Overall
Dataset
Method
Cox−nnet (dropout)*
Cox−nnet (ridge)
Cox−nnet (dropout+ridge)
Cox−PH (LASSO)
Cox−PH(ridge)*
Cox−PH (MCP)
Cox−nnet (two hidden layers)
Cox-nnet regularization method comparison
Cox-nnet (dropout + ridge)
Cox-nnet (ridge)
*Cox-nnet (dropout)
Rank
1
2
3
Cox-PH regularization method comparison
Cox-PH (MCP)
Cox-PH (LASSO)
*Cox-PH (Ridge)
Cox-nnet comparison between 1 and 2 hidden layers
*Cox-nnet (two hidden layers)
*Cox-nnet (one hidden layer)

Figure S3

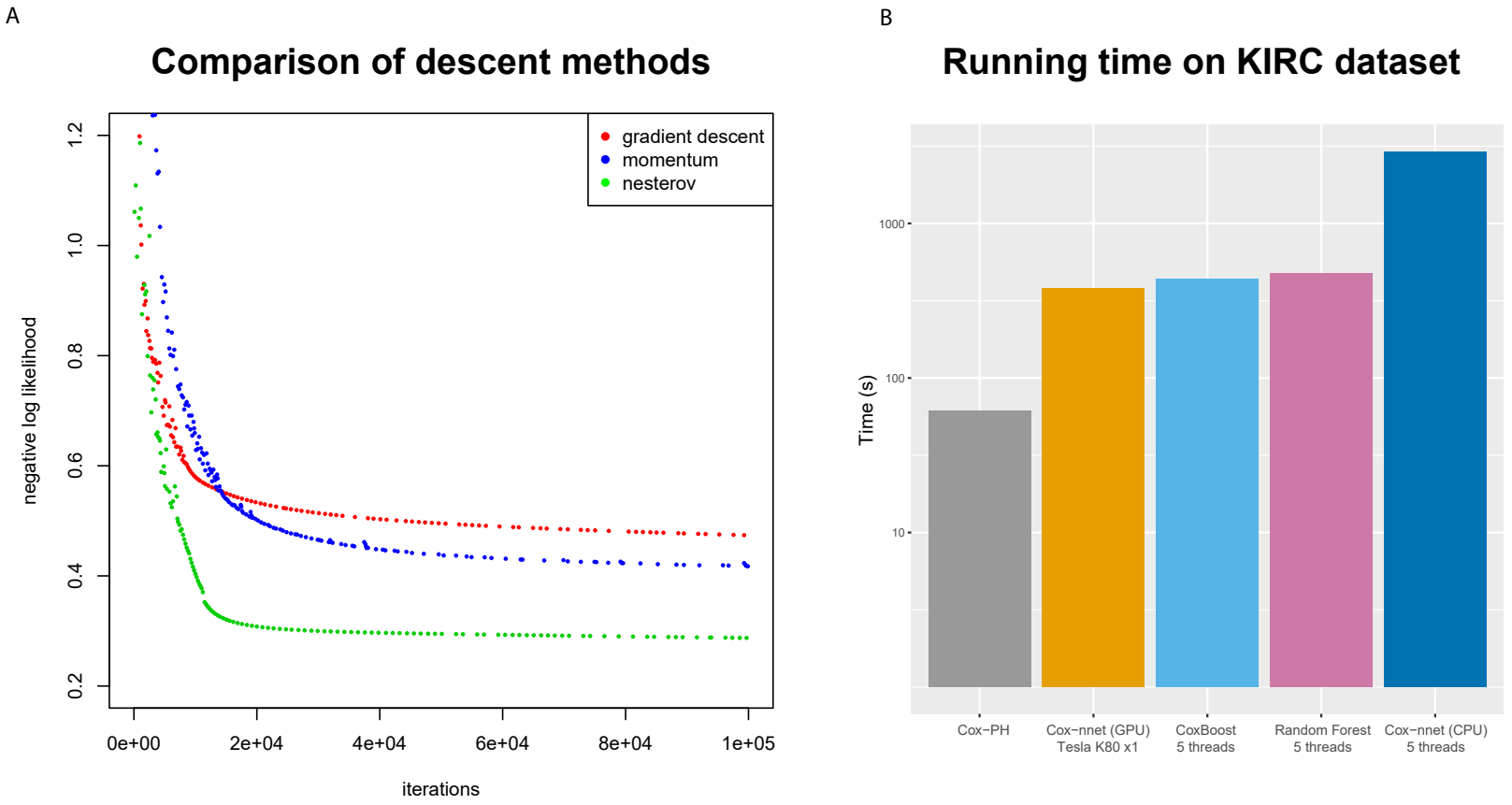

A
Comparison of descent methods
gradient descent
momentum
nesterov
negative log likelihood
0.2
0.4
0.6
0.8
1.0
1.2
0e+00
2e+04
4e+04
6e+04
8e+04
1e+05
iterations
B
Running time on KIRC dataset
Time (s)
1000
100
10
Cox−PH
Cox−nnet (GPU)
Tesla K80 x1
CoxBoost
5 threads
Random Forest
5 threads
Cox−nnet (CPU)
5 threads

Figure S4

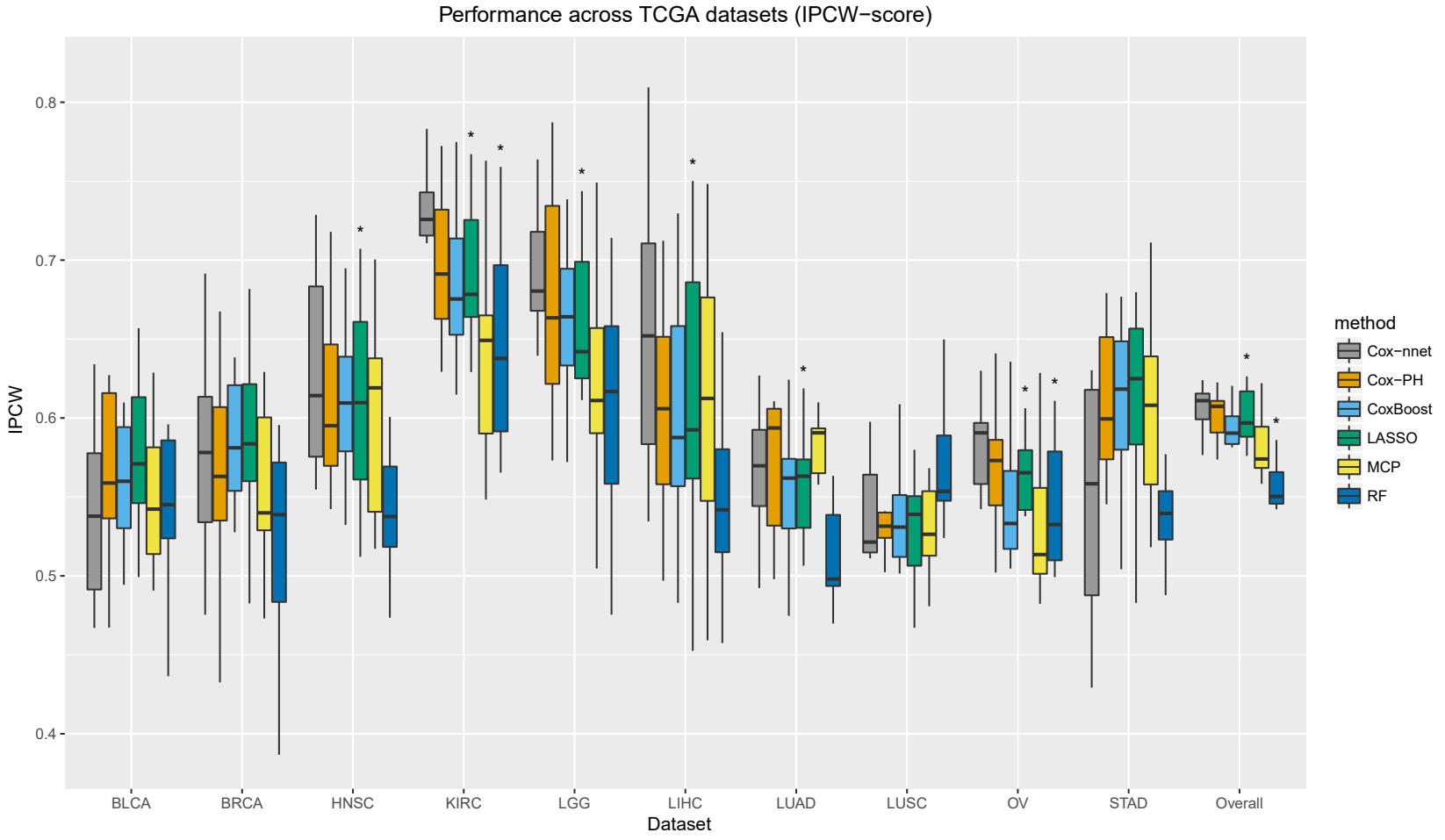

Performance across TCGA datasets (IPCW−score)
IPCW
0.8
0.7
0.6
0.5
0.4
BLCA
BRCA
HNSC
KIRC
LGG
LIHC
LUAD
LUSC
OV
STAD
Overall
Dataset
method
Cox−nnet
Cox−PH
CoxBoost
LASSO
MCP
RF

Figure S5

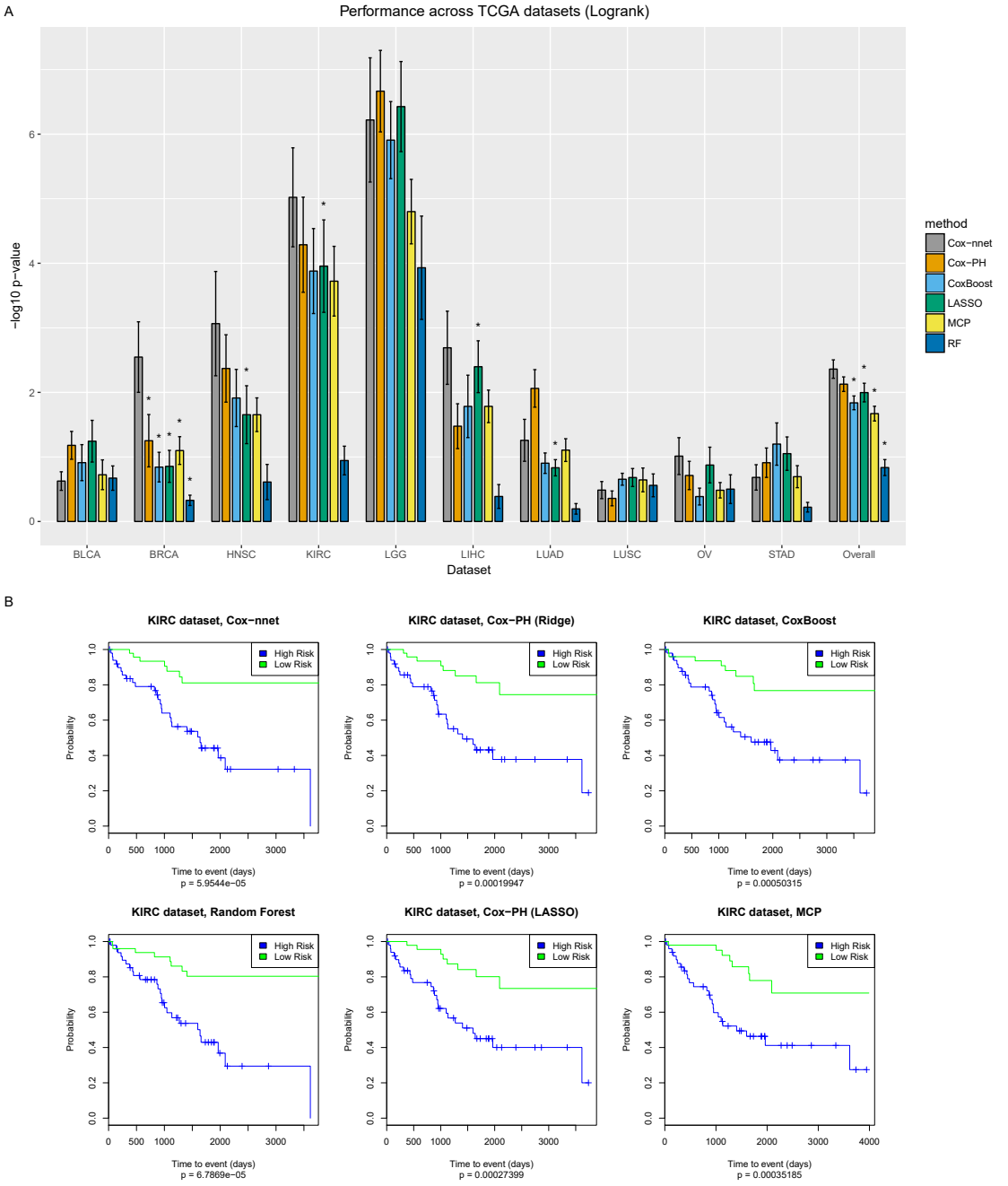

A
Performance across TCGA datasets (Logrank)
−log10 p−value
Dataset
BLCA
BRCA
HNSC
KIRC
LGG
LIHC
LUAD
LUSC
OV
STAD
Overall
method
Cox−nnet
Cox−PH
CoxBoost
LASSO
MCP
RF
B
KIRC dataset, Cox−nnet
p = 5.9544e−05
KIRC dataset, Cox−PH (Ridge)
p = 0.00019947
KIRC dataset, CoxBoost
p = 0.00050315
KIRC dataset, Random Forest
p = 6.7869e−05
KIRC dataset, Cox−PH (LASSO)
p = 0.00027399
KIRC dataset, MCP
p = 0.00035185
High Risk
Low Risk
Probability
Time to event (days)

Figure S6

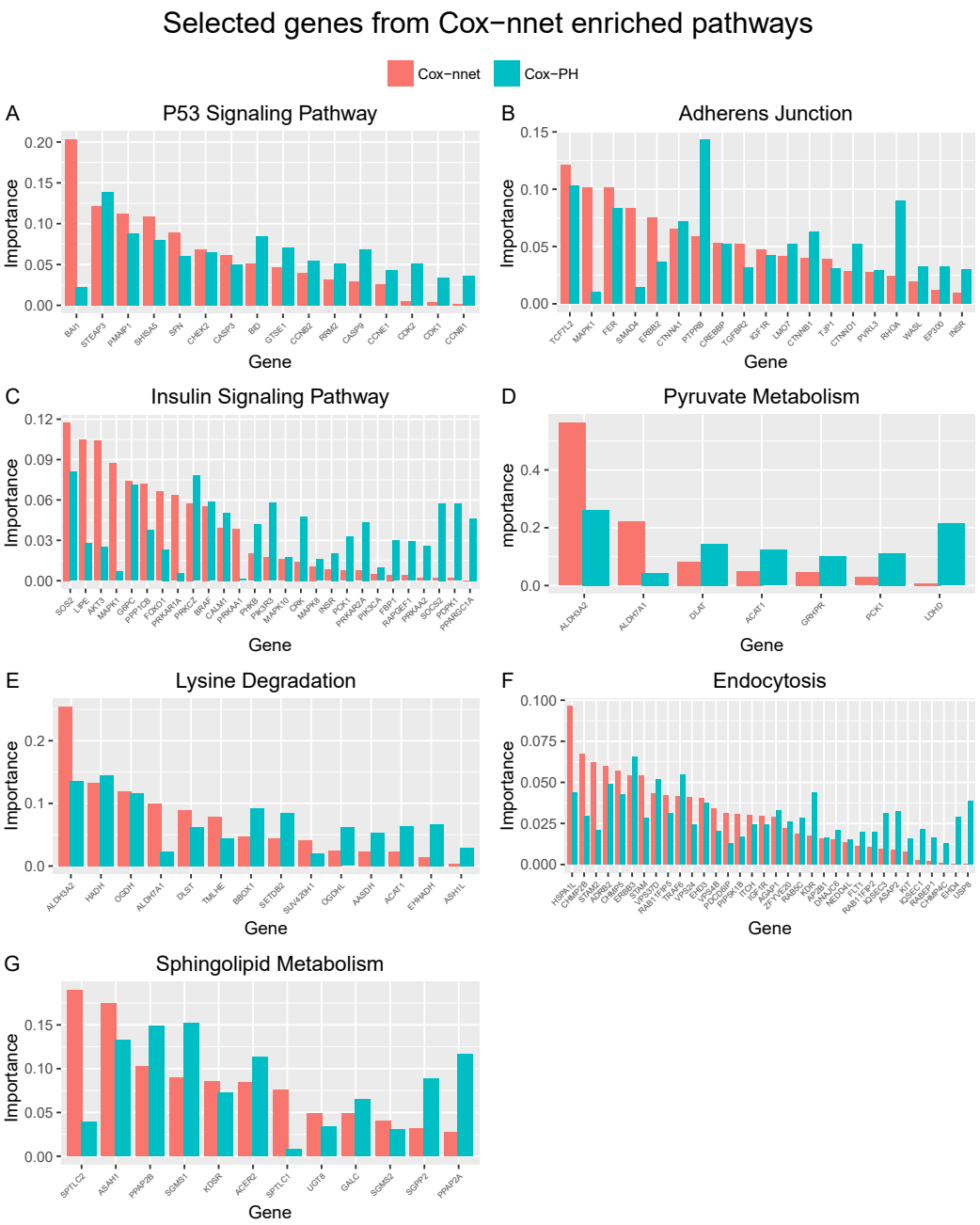
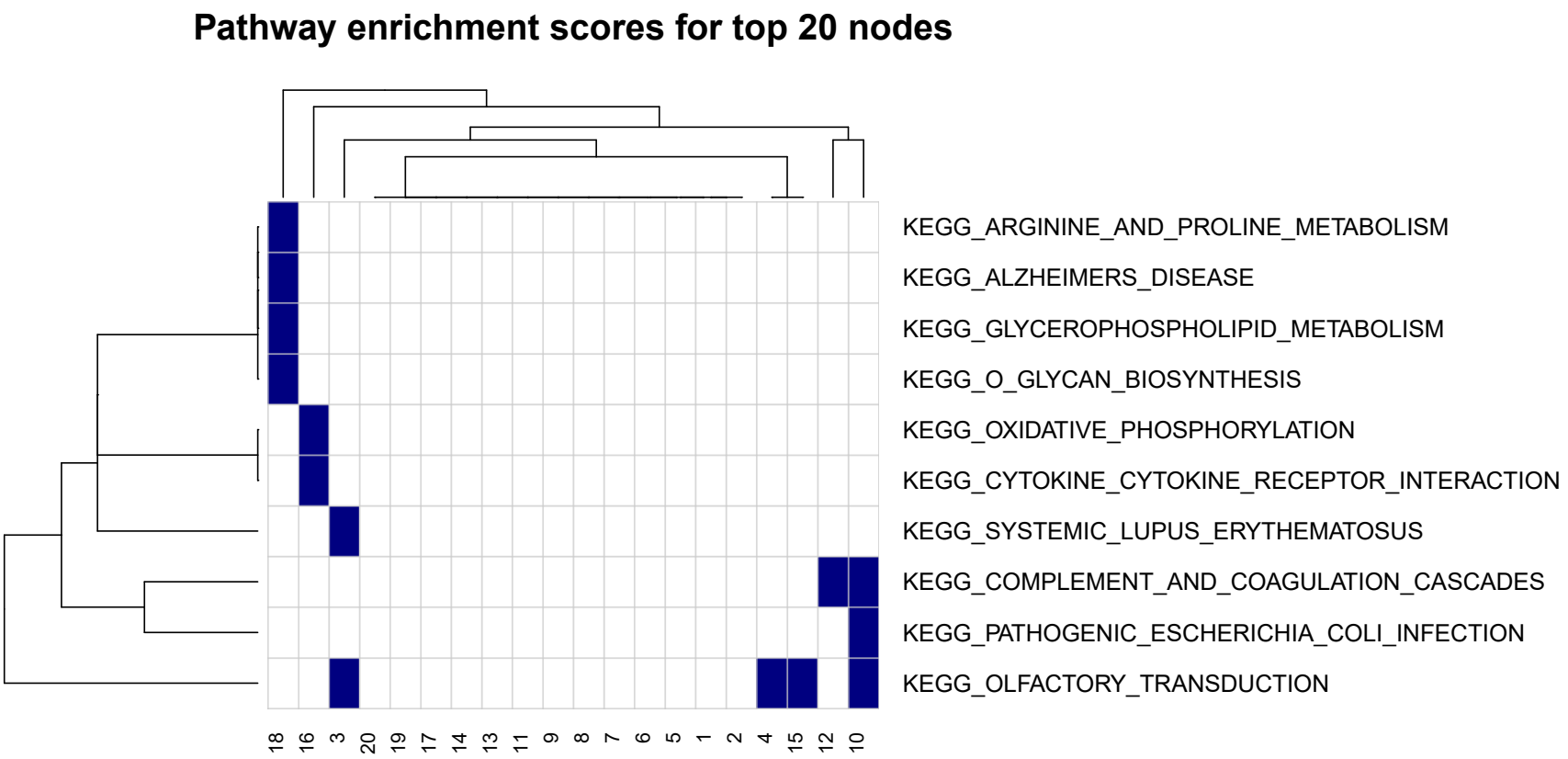

Selected genes from Cox−nnet enriched pathways
Cox−nnet
Cox−PH
A
P53 Signaling Pathway
B
Adherens Junction
C
Insulin Signaling Pathway
D
Pyruvate Metabolism
E
Lysine Degradation
F
Endocytosis
G
Sphingolipid Metabolism
Importance
Gene

Figure S7


Pathway enrichment scores for top 20 nodes
KEGG_ARGININE_AND_PROLINE_METABOLISM
KEGG_ALZHEIMERS_DISEASE
KEGG_GLYCEROPHOSPHOLIPID_METABOLISM
KEGG_O_GLYCAN_BIOSYNTHESIS
KEGG_OXIDATIVE_PHOSPHORYLATION
KEGG_CYTOKINE_CYTOKINE_RECEPTOR_INTERACTION
KEGG_SYSTEMIC_LUPUS_ERYTHEMATOSUS
KEGG_COMPLEMENT_AND_COAGULATION_CASCADES
KEGG_PATHOGENIC_ESCHERICHIA_COLI_INFECTION
KEGG_OLFACTORY_TRANSDUCTION
18 16 3 20 19 17 14 13 11 9 8 7 6 5 1 2 4 15 12 10

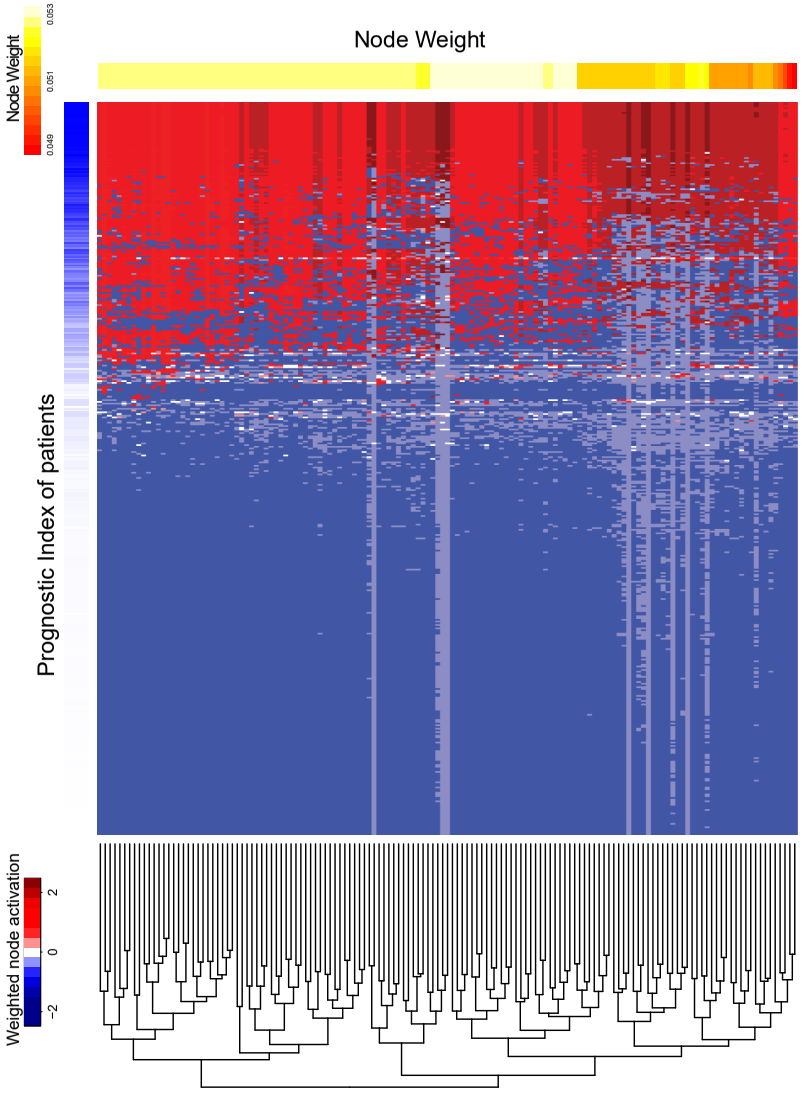

Node Weight
0.053
0.051
0.049
Node Weight
Prognostic Index of patients
Weighted node activation
2
0
−2

Figure S8

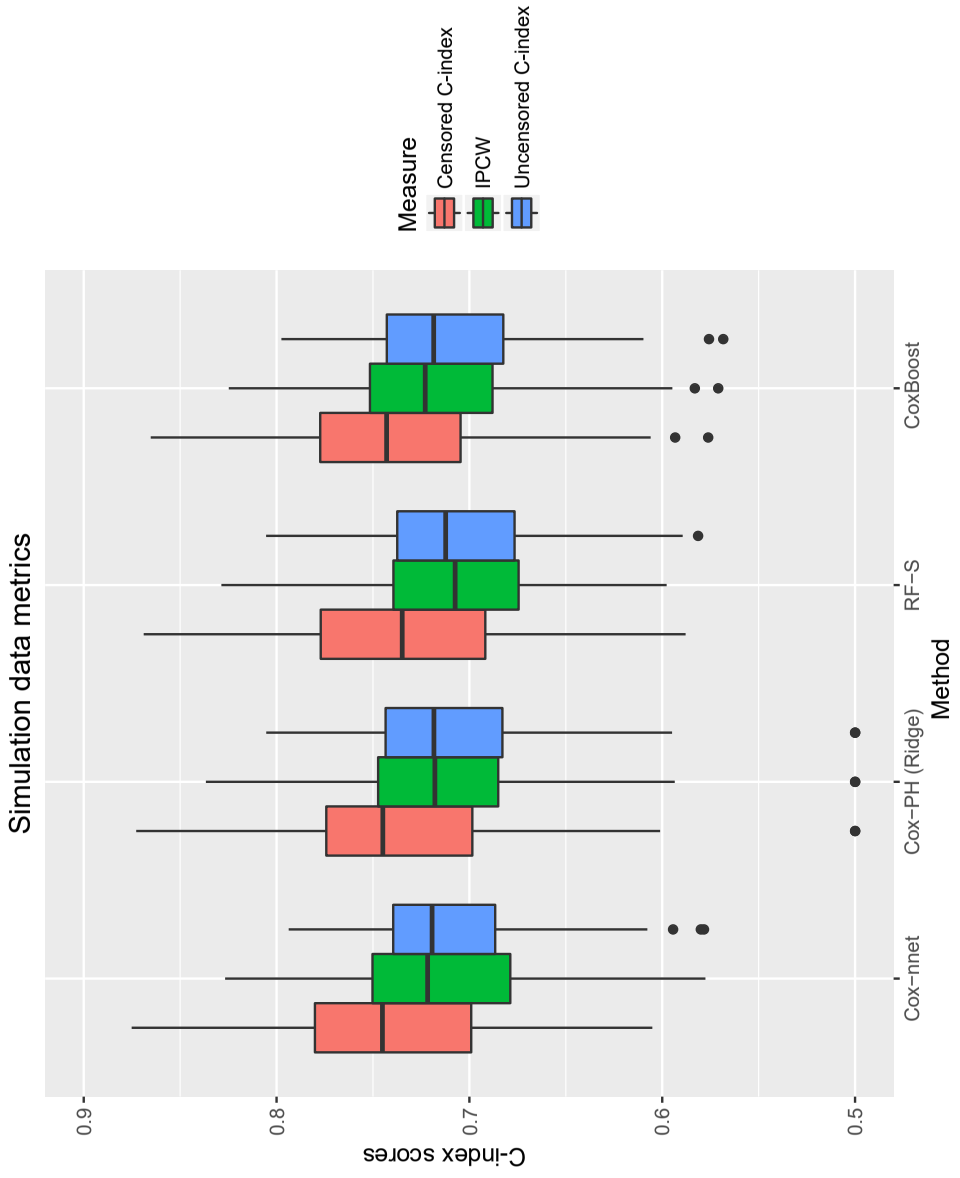

Measure
Censored C-index
IPCW
Uncensored C-index
Simulation data metrics
C-index scores
0.9
0.8
0.7
0.6
0.5
Cox−nnet
Cox−PH (Ridge)
RF−S
CoxBoost
Method

Figure S9

Table S1

Number of patient samples in TCGA datasets

Cancer Type	Event = 0	Event = 1	Total Samples
BLCA	229	177	406
BRCA	927	150	1077
HNSC	300	219	519
KIRC	356	175	531
LGG	391	121	512
LIHC	228	130	358
LUAD	324	166	490
LUSC	277	210	487
OV	130	172	302
STAD	215	134	349
Overall	3377	1654	5031

Table S2 part 1

Cox-nnet node-associated enriched pathways

Significantly enriched pathways common for all nodes ($p < 0.05$)

Pathway	P.value	P.adjusted	Nodes
KEGG adherens junction	0.000	0.001	1-20
KEGG butanoate metabolism	0.000	0.001	1-20
KEGG cell cycle	0.000	0.001	1-20
KEGG citrate cycle tca cycle	0.000	0.001	1-20
KEGG cytokine cytokine receptor interaction	0.000	0.001	1-20
KEGG endocytosis	0.000	0.001	1-20
KEGG endometrial cancer	0.000	0.001	1-20
KEGG fatty acid metabolism	0.000	0.001	1-20
KEGG insulin signaling pathway	0.000	0.001	1-20
KEGG lysine degradation	0.000	0.003	1-20
KEGG p53 signaling pathway	0.000	0.003	1-20
KEGG peroxisome	0.000	0.001	1-20
KEGG ppar signaling pathway	0.000	0.001	1-20
KEGG propanoate metabolism	0.000	0.001	1-20
KEGG prostate cancer	0.000	0.001	1-20
KEGG proteasome	0.000	0.001	1-20
KEGG proximal tubule bicarbonate reclamation	0.000	0.001	1-20
KEGG pyruvate metabolism	0.000	0.001	1-20
KEGG renin angiotensin system	0.000	0.002	1-20
KEGG sphingolipid metabolism	0.001	0.005	1-20
KEGG systemic lupus erythematosus	0.000	0.001	1-20
KEGG tight junction	0.000	0.001	1-20
KEGG tryptophan metabolism	0.000	0.001	1-20
KEGG valine leucine and isoleucine degradation	0.000	0.001	1-20
KEGG vascular smooth muscle contraction	0.000	0.001	1-20

Table S2 part 2

Additional significantly enriched pathways (p < 0.05)

Pathway	P.value	P.adjusted	Nodes
KEGG adipocytokine signaling pathway	0.001	0.005	1-19
KEGG beta alanine metabolism	0.000	0.003	1,3-20
KEGG colorectal cancer	0.000	0.001	1-13,15-20
KEGG homologous recombination	0.000	0.001	1-14,16-20
KEGG inositol phosphate metabolism	0.000	0.004	1-2,4-20
KEGG neurotrophin signaling pathway	0.000	0.001	1-13,15-20
KEGG nitrogen metabolism	0.001	0.007	1-16,18-20
KEGG phosphatidylinositol signaling system	0.000	0.003	1-2,4-20
KEGG aldosterone regulated sodium reabsorption	0.001	0.007	1-8,10-13,15-20
KEGG focal adhesion	0.000	0.001	1-13,16-20
KEGG mtor signaling pathway	0.002	0.012	1-5,7-13,15-20
KEGG renal cell carcinoma	0.000	0.001	1-13,15-18,20
KEGG ribosome	0.000	0.001	1-4,6-9,11-20
KEGG pathways in cancer	0.000	0.001	1-13,16-18,20
KEGG wnt signaling pathway	0.000	0.003	1-3,5-13,16-20
KEGG cytosolic dna sensing pathway	0.000	0.003	1-4,6-7,9,11-17,19-20
KEGG intestinal immune network for iga production	0.000	0.001	1-3,5,7-11,13-14,16-20
KEGG primary immunodeficiency	0.000	0.001	1-3,5,7-13,16-20
KEGG terpenoid backbone biosynthesis	0.002	0.011	1,3-5,7-13,16-20
KEGG tgf beta signaling pathway	0.000	0.001	1-13,17-18,20
KEGG vasopressin regulated water reabsorption	0.000	0.001	1-3,5,7-13,15,17-20
KEGG base excision repair	0.001	0.006	1-2,4-5,7-10,12-14,16-18
KEGG glycolysis gluconeogenesis	0.000	0.002	1,3-5,8,10-16,18-19
KEGG nod like receptor signaling pathway	0.000	0.001	2,5-7,9,11-16,19-20
KEGG erbb signaling pathway	0.000	0.003	1-4,8-12,17-18,20
KEGG pancreatic cancer	0.001	0.006	1-4,6,8,10-13,17-18
KEGG regulation of actin cytoskeleton	0.000	0.002	1-3,5-8,10,12-13,17-18
KEGG complement and coagulation cascades	0.000	0.001	4-5,7-9,11,14-16,19
KEGG dna replication	0.000	0.001	2,5,7,9,12-14,16,18-19

Table S2 part 3

Pathway	P.value	P.adjusted	Nodes
KEGG glycosaminoglycan biosynthesis chondroitin sulfate	0.001	0.005	2,4,6,9,14-17,19-20
KEGG non small cell lung cancer	0.002	0.012	1,3-4,6,8,10-11,15,17
KEGG pathogenic escherichia coli infection	0.000	0.001	2,4,6-7,9,14-16,19
KEGG chronic myeloid leukemia	0.001	0.009	1,3,8,10-12,17-18
KEGG pyrimidine metabolism	0.000	0.004	2,4,6-9,14,17
KEGG thyroid cancer	0.006	0.023	1,3,5,8,10,12,17-18
KEGG graft versus host disease	0.000	0.001	7-9,11,16,19-20
KEGG leukocyte transendothelial migration	0.003	0.014	1,3,5,10,12-13,18
KEGG melanoma	0.005	0.023	1,3,8,10,12,17-18
KEGG arginine and proline metabolism	0.004	0.019	1,3-4,8,10,12
KEGG histidine metabolism	0.004	0.019	3-4,9,12,14,19
KEGG oxidative phosphorylation	0.000	0.001	1,5,10,12-13,19
KEGG long term potentiation	0.004	0.020	1,10,12,17-18
KEGG taste transduction	0.005	0.021	1-3,5,12
KEGG acute myeloid leukemia	0.012	0.041	4,8,10,17
KEGG allograft rejection	0.000	0.001	8-9,16,19
KEGG arrhythmogenic right ventricular cardiomyopathy arvc	0.011	0.038	2-3,7,18
KEGG glioma	0.006	0.025	1,8,10,17
KEGG leishmania infection	0.000	0.001	9,14,16,19
KEGG lysosome	0.000	0.002	1,3,10,12
KEGG natural killer cell mediated cytotoxicity	0.000	0.001	7,9,16,19
KEGG o glycan biosynthesis	0.008	0.031	1-2,6,17
KEGG parkinsons disease	0.005	0.020	1,5,10,12
KEGG spliceosome	0.001	0.007	6,11,14,17
KEGG calcium signaling pathway	0.003	0.015	1,12,18
KEGG gap junction	0.007	0.028	1,17-18
KEGG hematopoietic cell lineage	0.000	0.001	9,16,19
KEGG small cell lung cancer	0.005	0.022	8,10,17
KEGG toll like receptor signaling pathway	0.000	0.002	9,16,19
KEGG type i diabetes mellitus	0.000	0.001	9,16,19

Table S2 part 4

Pathway	P.value	P.adjusted	Nodes
KEGG ubiquitin mediated proteolysis	0.006	0.026	1,10,17
KEGG amyotrophic lateral sclerosis als	0.012	0.041	6,19
KEGG antigen processing and presentation	0.000	0.001	16,19
KEGG autoimmune thyroid disease	0.000	0.002	16,19
KEGG chemokine signaling pathway	0.000	0.001	16,19
KEGG fc epsilon ri signaling pathway	0.013	0.043	4,17
KEGG huntingtons disease	0.004	0.020	1,19
KEGG jak stat signaling pathway	0.002	0.011	16,19
KEGG long term depression	0.006	0.024	1,17
KEGG melanogenesis	0.012	0.040	10,18
KEGG prion diseases	0.006	0.025	14,19
KEGG alzheimers disease	0.003	0.013	1
KEGG ascorbate and aldarate metabolism	0.003	0.016	4
KEGG asthma	0.003	0.016	19
KEGG basal cell carcinoma	0.006	0.023	4
KEGG cell adhesion molecules cams	0.006	0.024	19
KEGG drug metabolism cytochrome p450	0.012	0.040	4
KEGG fc gamma r mediated phagocytosis	0.004	0.017	19
KEGG glycine serine and threonine metabolism	0.009	0.033	4
KEGG glycosaminoglycan biosynthesis keratan sulfate	0.010	0.035	14
KEGG mismatch repair	0.016	0.050	7
KEGG primary bile acid biosynthesis	0.007	0.026	5
KEGG purine metabolism	0.012	0.041	14
KEGG steroid biosynthesis	0.008	0.029	19
KEGG t cell receptor signaling pathway	0.004	0.017	19
KEGG viral myocarditis	0.000	0.003	19

Table S3 part 1

GSEA enrichment of Cox-PH associated cancer pathways

List of significantly enriched pathways (p < 0.05)

Pathway	E.score	P.value	P.adjusted
KEGG cell cycle	-0.407	0.000	0.003
KEGG citrate cycle tca cycle	0.607	0.000	0.003
KEGG fatty acid metabolism	0.561	0.000	0.003
KEGG lysosome	0.410	0.000	0.003
KEGG oxidative phosphorylation	0.418	0.000	0.003
KEGG proximal tubule bicarbonate reclamation	0.633	0.000	0.003
KEGG valine leucine and isoleucine degradation	0.577	0.000	0.003
KEGG olfactory transduction	0.294	0.000	0.005
KEGG cytokine cytokine receptor interaction	-0.330	0.000	0.005
KEGG proteasome	-0.540	0.000	0.005
KEGG vascular smooth muscle contraction	0.391	0.000	0.005
KEGG peroxisome	0.415	0.000	0.006
KEGG tight junction	0.343	0.000	0.006
KEGG propanoate metabolism	0.518	0.001	0.009
KEGG ppar signaling pathway	0.406	0.001	0.011
KEGG renin angiotensin system	0.645	0.001	0.011
KEGG systemic lupus erythematosus	-0.377	0.001	0.014
KEGG prostate cancer	0.368	0.002	0.016
KEGG pathogenic escherichia coli infection	-0.433	0.003	0.034
KEGG endometrial cancer	0.409	0.004	0.040
KEGG butanoate metabolism	0.458	0.005	0.042
KEGG epithelial cell signaling in helicobacter pylori infection	0.361	0.006	0.047
KEGG nod like receptor signaling pathway	-0.407	0.006	0.047
KEGG beta alanine metabolism	0.521	0.006	0.047
KEGG complement and coagulation cascades	-0.393	0.007	0.047

Table S3 part 2

Pathway	E.score	P.value	P.adjusted
KEGG tryptophan metabolism	0.427	0.007	0.047
KEGG arginine and proline metabolism	0.381	0.007	0.047
KEGG calcium signaling pathway	0.272	0.007	0.047
KEGG spliceosome	-0.333	0.007	0.047
KEGG parkinsons disease	0.304	0.008	0.048

5.9 Chapter summary

In this chapter, we further explore the application of machine learning methods to high throughput RNA-Seq data. We used artificial neural networks (ANN) in order to predict cancer patient survival. ANN is a machine learning method that is the basis for the deep innovation of self-driving cars. Here, we create a new model, termed Cox-nnet, applying ANN to censored survival data. Unlike the standard Cox regression method, this approach may pick up on interactions between features (in this case, gene expression) to more accurately predict patient survival. In addition, we show that information extracted from the hidden layer can be used to reveal useful biological pathway information from each patient.

Chapter 6

Pan-cancer analysis of expressed single nucleotide variants in long intergenic non-coding RNA

Travers Ching[1,2], Lana X. Garmire[1,2]

Manuscript in preparation.

[1]University of Hawaii Cancer Center, 701 Ilalo Street, Honolulu, Hawaii, USA 96813

[2]Graduate Program of Molecular Biosciences and Bioengineering, University of Hawaii-Manoa, 1955 East-West Road, Honolulu, Hawaii, USA 96822

6.1 Preface

Long intergenic non-coding RNAs (lincRNAs) are emerging as important components in cancer biology. It is understood that changes and mutations in cancer driver genes effect a series of downstream expression changes. However, because lincRNAs are a relatively new class of transcripts compared to protein coding genes, the mutational landscape of lincRNAs has not been as extensively studied.

Although RNA-Seq data are most commonly used to determine gene expression and analyze alternative splicing events, a number of studies have shown that robust variant calling can be

performed using RNA-Seq data. In the current study, we determined expressed lincRNA and protein coding gene variations on primary tumors from 6000 patients in 12 different cancer types. We highlight the relationship of molecular and genetic features related to lincRNA mutations in cancer and explore the relationship of eSNVs and determine regions in the genome that have a high frequency of lincRNA mutations.

Exploring the landscape of lincRNA somatic mutations in cancer may eventually lead to a more fundamental understanding of the tumorigenic process of genetic mutations leading to malignancy.

6.2 Introduction

Long intergenic non-coding RNAs (lincRNAs) are emerging as important components in cancer biology. The expression of thousands of lincRNAs are dysregulated in cancer, and many lincRNAs have been shown to be robust biomarkers for tumor tissue and patient prognosis [1–3]. There is also strong evidence that lincRNAs may serve as drivers of tumorigenesis, drug resistance and disease progression [4–6]. It is understood that changes and mutations in cancer driver genes effect a series of downstream expression changes [7, 8]. However, because lincRNAs are a relatively new class of transcripts compared to protein coding genes, the mutational landscape of lincRNAs has not been as extensively studied.

Differences in single nucleotide positions from the reference genome may arise through genetic inheritance (germline) or occur spontaneously in the genome of cells in the body (somatic). Germline nucleotide differences that are shared between members of a population are termed Single Nucleotide Polymorphisms (SNPs) [9]. In the context of cancer, somatic single nucleotide changes not found in the germline are termed Single Nucleotide Variants (SNVs) [9]. When SNVs are found on expressed genes or transcripts in the transcriptome, they are termed expressed SNVs (eSNVs) [10].

Exome sequencing is a popular platform that has been extensively used to investigate genetic differences and mutations in cancer. Exome sequencing captures targeted DNA through PCR or other enrichment strategies, and then uses next generation sequencing platforms to determine the nucleotide sequences of the targeted regions [11]. However, exome sequencing traditionally interrogates only the exons of protein coding regions, and does not include the majority of non-coding transcripts [12]. Whole genome sequencing (WGS) could potentially provide the same information and more. However, the number of WGS sample data are far more limited compared

to exome sequencing and RNA-Seq data, and also do not contain expression information and expressed allele information.

It has been emphasized that extensive heterogeneity exists between tumor samples as well as within tumor samples. Furthermore, cellular subpopulations are mixed in tumor samples. Not only is the expression of a gene important, but the particular isoforms and gene alleles may be essential in characterizing the tumor heterogeneity and cancer pathways. SNVs may influence the expression of proximal genes (in cis) or genes far away or downstream in a molecular pathway (in trans) [13, 14]. Thus, to interrogate the effects of lincRNA mutations, we used RNA-Seq data to perform variant calling on both lincRNA associated regions and protein coding genes.

Although RNA-Seq data are most commonly used to determine gene expression and analyze alternative splicing events, a number of studies have shown that robust variant calling can be performed using RNA-Seq data [10, 15, 16]. Using The Cancer Genome Atlas (TCGA) RNA-Seq data, we determine expressed lincRNA and protein coding gene variations on primary tumors from over 6000 patients in 12 different cancer types. We highlight the relationship of molecular and genetic features related to lincRNA mutations in cancer and explore the relationship of eSNVs present in lincRNA transcribed regions with their gene expression. Finally, we compare the molecular features strongly correlated with lincRNA mutations and those in protein-coding genes.

6.3 Methods

6.3.1 TCGA Datasets

We used 12 cancer datasets from TCGA and included 6118 primary tumor samples in this study. These datasets include bladder urothelial carcinoma (BLCA) breast invasive carcinoma (BRCA), head and neck squamous cell carcinoma (HSNC), kidney renal clear cell carcinoma (KIRC), Liver hepatocellular carcinoma (LIHC), low grade glioma (LGG), lung adenocarcinoma (LUAD), lung squamous cell carcinoma (LUSC), Ovarian serous cystadenocarcinoma (OV), and stomach adenocarcinoma (STAD), prostate adenocarcinoma (PRAD) and thyroid carcinoma (THCA). Both RNA-Seq BAM files and RNA-Seq fastq files were downloaded using the GeneTorrent program from the UCSC Cancer Genomics Hub (`https://cghub.ucsc.edu`). Additional TCGA samples were downloaded from NCBI Genomic Data Commons Data Portal (`https://gdc-portal.nci.nih.gov`) using the GDC data transfer tool. In total, 6118 samples with both expression data and eSNV data were used in the analysis (Table S1 and S2).

6.3.2 Expresion quantification

LincRNA were quantified from GRCh37/hg19 aligned BAM files from TCGA using standard procedures [2]. The FeatureCounts program [17] was used to calculate raw counts for each sample, based on lncipedia 3.1 annotation. Intergenic transcripts from lncipedia were determined by subtracting transcripts that overlapped with the transcription coordinates of protein coding genes from ENCODE [18]. Count data was normalized using the DESeq2 package in R [19] and fragments per kilo bases of exons per million mapped reads (FPKM) were estimated (See Figure S1).

For protein coding genes, raw counts were downloaded from the Broad institute firehose project (`http://gdac.broadinstitute.org`). Protein coding genes were determined using the BiomaRt R package [20], using the "gene_biotype" field. Count data was normalized and FPKM was estimated using the DESeq2 package.

6.3.3 Exome sequencing comparison

The exome sequencing variant calls, including somatic and germline variants, were downloaded for 7 TCGA datasets (BLCA, HNSC, KIRC, LGG, LIHC, LUAD, PRAD and STAD). For comparison, eSNVs from the RNA-Seq dataset were filtered by the probe coverage region of the exome sequencing data. The proportion of eSNVs also called by the exome sequencing variant calls in each sample was used as a quality measure for the correctness of the eSNV variant calling procedures.

6.3.4 Predicting germline and somatic mutations

Using the exome sequencing somatic and germline mutation data from TCGA, we built a Random Forest model to classify somatic vs. germline SNVs. The Xgboost package in R was used (version 0.6-0) with 1000 trees. Four features were used in the building of this model: dbsnp (whether an SNV occurred at a position annotated by the dbSNP database), fa.tumor (the estimate allele ratio of the SNV in the tumor exome sample), conservation (PhyloP conservation score from the UCSC genome browser) and transversion (whether the SNV was a transition or transversion mutation). The data was split into 80 traning and 20% testing for evaluation.

6.3.5 Expressed single nucleotide variations (eSNVs)

Raw read data was downloaded from UCSC Cancer Genomics Hub in fastq format. Reads were first aligned to the hg19 genome reference using STAR aligner [21] in two-pass mode. Aligned BAM files were sorted using Picard-tools' ReorderSam function (`http://broadinstitute.github.io/picard`) and reads were split based on splicing junctions using Genome Analysis Toolkit (GATK) SplitNCigarReads function [22]. Reads were then processed through removal of duplicates, indel realignment and base recalibration following standard protocols. Variant calling was performed using GATK's Haplotype caller. Data processing was performed on the University of Hawaii high performance computing cluster (see Supplementary Methods for details). To further reduce potential false positive calls, variants were filtered based on SNV clusters and read strand bias following recommendations from the developers. To define lincRNA specific eSNVs, variants associated with lincRNAs based on the lncipedia 4.0 reference [4] were saved for analysis. eSNVs for protein coding genes were similarly filtered based on the Refseq transcriptome annotation.

6.3.6 Predictive models for eSNVs

We constructed classification models in order to predict eSNVs for each cancer type. These models were built on balanced datasets, comprised of somatic eSNVs and background "negative" eSNVs – i.e., random sites on expressed lincRNAs in each RNA-Seq sample. The molecular features used in these models included conservation, copy number variation, histone marker features, nucleotide composition features, etc. (Table S3). Two algorithms were employed on these datasets: logistic regression with ridge regression (LR), a fast linear classification algorithm using the glmnet R package (version 2.0-5) and Gradient Boosted Trees [23], a fast non-linear tree-based classifier using the xgboost R package (version 0.6-0).

To evaluate each model, the datasets were split into 80% training and 20% testing. AUC was calculated as performance metrics on the testing sets. To evaluate the relative importance of each feature, we used the Gain measurement, which is the improvement of the model accuracy on its branches in the trees for each feature.

6.3.7 Calculating mutation probabilities

Although the Boosted Trees models were built using balanced datasets, the true negative class of the eSNV data are all mutation sites that were not called as variants. The posterior probabilities

of the Boosted Trees models must be modified to account for the increase in type I errors that would result from application of the models to the entire lincRNA transcriptome. In order to calculate individual site mutation probabilities, we adjusted the posterior probabilities of the Boosted Trees eSNV models through a Bayesian framework [24]. For each patient with RNA-Seq data in this study, we calculated the mutation probabilities for every genomic position included in a lincRNA. Subsequently, we calculated the log-odds mutation scores.

6.3.8 Calculating feature importance and feature mutual information

In order to determine feature importance, we calculate the Gain value of each feature. In an ensemble forest model (Random Forest or Gradient Boosted Trees), Gain is the average improvement of performance of the model on each tree branch that is split by the feature in the ensemble forest [23].

To determine the correlation between features, we calculated the normalized mutual information [25]. Mutual information is a measure of correlation between two variables [26]. It measures the information shared between two variables based on their conditional probability distributions. The normalized mutual information is a measure that scales mutual information by the geometric mean of geometric mean of the two variables, and therefore ranges between 0 and 1.

6.4 Results

6.4.1 Computational pipeline accurately predicts genetic variation in tumor RNA-Seq samples

6118 primary tumor RNA-Seq samples were selected from 12 TCGA datasets (Table S1). A pipeline for calling mutations from bulk RNA-Seq data was implemented (Figure S1). TCGA fastq files were aligned using STAR aligner, and then pre-processed by splitting reads along exon junctions. These samples were then further pre-processed using the Indel Realignment and Base Recalibration modules in GATK. Finally, VCF files were generated using the Haplotype caller GATK module. To verify the quality of the results, we compared the variant calls from exome sequencing for paired exome and RNA-Seq sample datasets (Figure S2). On average, 80% of the single nucleotide variants found in RNA-Seq data were also found in the exome sequencing variant calls, inside the regions of the exome sequencing probes (Figure S2 and S3). O'Rawe et al. 2013 found that the concordance between sequencing platforms and variant calling software to be about 50% [27]. Furthermore, the sensitivity of state of the art variant calling platforms

is about 55% [28]. Thus, the high proportion of eSNVs detected in both RNA-Seq and exome platforms suggests the variant calls from the RNA-Seq are reliable. Furthermore, we compared the concordance between exome and RNA-Seq samples in each dataset (Figure S3). We found that the paired Exome and RNA-Seq samples had much higher concordant SNVs compared to samples from different patients.

6.4.2 A Random Forest model differentiates somatic and germline mutations

Using the exome sequencing data, we built a random forest model classifying the somatic mutations versus the germline mutations based upon four features: dbsnp (whether the mutation is documented in the NCBI dbSNP database), FA.tumor (the fraction of the alternate allele in the tumor sample), conservation (PhyloP conservation score) and transversion (whether the mutation is a transversion or a transition mutation). This model had an AUC of 0.988 on the exome sequencing data and an AUC of 0.983 and 0.987 on based on the exome and RNA-Seq data respectively (Figure 1A). By comparison, the logistic model had slightly lower AUCs of 0.979 and 0.985. The dbsnp and FA.tumor features had relatively high importance scores in correlation with the outcome (somatic or germline mutation), whereas conservation and transversion were not important features in this model (Figure 1B).

Secondly, we then applied this model to the 12 RNA-Seq datasets, and selected eSNVs which were highly confident of being germline (posterior probability < 0.05) or somatic mutations (posterior probability > 0.95). Using these thresholds, 170 million germline variants were detected (155.4 million in protein-coding genes and 14.56 million in lincRNA genes) and 5.67 million eSNVs were detected (5 million in protein coding genes and 660,000 in lincRNA genes) (Table S1). Within lincRNA regions, there 2.48 million somatic mutations and 19.3 million germline mutations. Within protein coding exonic gene regions, there were 9.46 million somatic mutations and 136.9 million germline mutations.

6.4.3 lincRNA eSNV genome-wide landscape

To explore lincRNA somatic mutations, we plotted the density of eSNVs by binning lincRNA eSNVs in 100,000 base pair windows across the genome, and then normalizing by the exon density of the lincRNA transcriptome (Figure 2). There are some regions that have an increased frequency of lincRNA eSNVs. The top four regions included chr2p11.2, chr14q32.33, chr7q32.1 and chr1p36.13. In particular, chr2p11.2 is known to be heavily associated with breast cancer [29]. In 2011, Sahin et al. found that copy number inbalances in chr2p11.2 had a significant effect

on breast cancers detected through symptoms vs. breast cancers detected through screening [29]. They also found that the imbalance had a significant effect on disease free survival. However, they were not able to determine any association with protein coding genes. These results suggest that the association of this region with cancer phenotypes could be due to lincRNA mutations.

6.4.4 A gradient boosted model determines eSNV mutation likelihood

For each of the 12 TCGA cancer types, we constructed a classification model to predict mutation likelihood. We constructed balanced dataset comprised of somatic eSNVs (positive class) and negative "background" eSNV; e.g., random positions on expressed genes or lincRNAs which are not mutated. We extracted nucleotide, position, gene level and tissue features relevant to each individual eSNV (descriptions of features used are in Table S2). In addition, for eSNVs located within protein coding genes, we extracted features related to the coding sequence frame. Similarly, we also built models using the germline variants using the same features.

We applied three machine learning algorithms to each dataset: logistic regression (a linear classifier), a neural networks (a flexible non-linear classifier) and gradient boosted trees (a fast tree-based non-linear classifier). In each dataset, the Boosted Trees model performed considerably better than the neural network and logistic regression models. The neural network models generally performed better than the logistic regression. Across all 12 TCGA datasets, Boosted Trees had an average AUC of approximately 0.947 (Figure 3). The neural network and logistic regression models had average AUCs of 0.876 0.839, respectively in differentiating somatic eSNVs in lincRNA (Figure S4). For somatic mutations in protein coding genes, Boosted Trees had an average AUC of 0.930; neural network logistic regression had average AUCs of 0.908 and 0.845, respectively (Figure 3).

By comparison, the Boosted Trees models built on predicting germline variants had AUCs of 0.886 (lincRNAs) and 0.883 (protein coding genes) (Figure S4). Similar to the eSNV models, the neural network and logistic regression models for germline variants had significantly lower AUCs of 0.828 and 0.772 for lincRNAs; and 0.883 and 0.815 for protein coding genes. In all cases, the Boosted Trees models had stronger performance compared to logistic regression by large margins. This suggests that the relationship of the features to the outcome are not linear and may be complex and not easily understood.

6.4.5 Molecular features correlating with somatic eSNVs differ from germline variants and differ from protein coding genes

To evaluate the importance of each feature in the model, we used the Gain measure, which calculates the average increase in performance for each feature in every tree in the Boosted Trees ensemble. For the lincRNA eSNV models (Figure 4), transversion (whether the SNV was the result of a transversion mutation), followed by conservation. For the two lung cancer datasets (LUAD and LUSC) the histone H3k09me3 methylation feature from a lung fibroblast cell line, Ag04450H3k09me3_pos, was the third most important feature. For other datasets, CG_0, the nucleotide base feature, and cnv_pos (the copy number variation at the SNV position) were the next most important features. Interestingly, for protein coding gene mutations, the UTR3 feature, the location of a eSNV on the 3'UTR region, was the third most important feature, and the CG_0 and cnv_pos features were relatively less important.

For germline lincRNA variants (Figure S5), conservation was the most important feature, while transverison was the 2nd most important. Cnv_pos did not come up as an important feature. Interestingly, several histone features were important in specific datasets. For kidney renal cell carcinoma, LncapH3k04me3_pos and Hepg2H3k4me3_pos (the trimethylation histone signature for a prostate and liver cancer cell line respectively) was the third and fourth most important features. Promoter methylation signatures were relatively less important than methylation signatures at the eSNV position. In addition, nucleotide composition upstream or downstream of the eSNV were not as important as the nucleotide base of the eSNV position. Only the nucleotide base at the eSNV position were determined to be important, with C/G nucleotides being much less likely to be mutated (Figure S6). While cnv_pos was an important feature, there was not much separation between the distributions of the positive and negative classes in the models (Figure S7), suggesting that the role of copy number variation is complex.

For protein coding somatic eSNVs, features related to the protein coding frame were also included. Similar to the model for lincRNA eSNVs, transversion was the most important feature. However, utr3 (whether a mutation occurred in the 3'UTR region) was the 3rd most important feature. Cnv_pos (copy number variation at the eSNV position) was less important for protein coding genes.

6.4.6 Feature correlation is determined through normalized mutual information

To explore how features are related to one another, we calculated the normalized mutual information for features used to predict somatic eSNVs (Figure 5) and performed hierarchical clustering on the pairwise mutual information scores. There were three distinct clusters from the histone data. H3k4 trimethylation histone features were split based on the location of measurement (i.e., at the histone values at the promoter site or the nucleotide position). Interestingly, the features in these two groups were only weakly correlated. Other types of methylation (H3k27, H3k36) weakly clustered together, although the correlation between these features were much weaker. The cnv_pos and cnv_promoter features were almost identical, as copy number variation was defined for large sections of the genome. Otherwise, no other feature had strong correlation with other features, suggesting that they had orthogonal information. Interestingly, lincRNA gene expression ("expr") did not strongly correlate with any feature, and only weakly correlated with histone features at the promoter regions.

6.4.7 LincRNA tumor drivers have distinct mutation profiles

To explore the mutation landscape of lincRNAs, we calculated a summarized mutation score as the average mutation log odds for every nucleotide position in all lincRNAs. We converted this probability to log odds and found the mean log odds for every lincRNA in each TCGA sample. Next we asked whether known driver lincRNAs had a different mutation profile compared non-driver lincRNAs. We used the lnc2Cancer database of experimentally validated lncRNA drivers [30], and found approximately 500 intergenic lincRNAs in the database. The known driver lincRNAs showed distinct mutation and expression profiles in several cancer types (Figure 6). Overall known driver lincRNAs had significantly lower mean log odds compared to non-drivers ($p < 2.2e\text{-}16$) and higher overall expression ($p < 2.2e\text{-}16$) (Figure S8).

6.4.8 LincRNA eSNV profiles provides more robust clustering compared to lincRNA expression

We hypothesized that since the eSNV data incorporates RNA-Seq, copy number variation, histone data, etc., whereas the expression data only contains RNA-Seq information, the eSNV data would provide better clustering results. Here we performed sample link community clustering [31] on the eSNV data (using summarized mutation scores) and compared the results to the clustering based on the lincRNA expression (Figure 7). The Link Community clustering

method purports to find robust communities on complex and overlapping clusters, based on similarity of node edges. On average, the eSNV clustering had much stronger robustness, as measured by the partition density, compared to the clustering determined by expression (Figure S9). This suggests that the clusters determined by eSNV data may have a stronger signal than the clusters determined by expression data.

6.5 Discussion

As others have previously investigated, it is possible to accurately determine somatic mutations in cancer samples without paired normal samples [32]. In this study, we show that a simple model can accurately separate eSNVs from germline SNPs, based on models built on exome sequencing data and tested on RNA-Seq data. Since paired normal samples are often not available, as in the case of older FFPE DNA samples [33], or RNA-Seq expression samples in the present case, it is therefore useful to be able to computationally predict which variants are somatic mutations. Based on the random forest model (Figure 3), the features dbsnp (whether a variant is found in the dbSNP database) and allele frequency are important features in predicting which variants are somatic. dbSNP variants are those commonly found in population germlines, and therefore are much less likely to be somatic. Similarly, it has been noted that variants that have low allele frequency are likely to be cancer mutations, and may even play important roles in cancer development [34]. Thus, variants found to have 100% allele frequency in the tumor samples are unlikely to be somatic mutations, as normal sample contamination is usually present [35]. Furthermore, even in normal sample contamination is removed, tumor samples often contain multiple populations that may have different alleles and mutational profiles [36].

The models predicting eSNVs from the background nucleotide positions showed strong performance (Figure 3), showing that it is possible to differentiate sites which are eSNVs from those which are unlikely to be eSNVs. Comparing the different classification algorithms, the logistic regression models for protein-coding gene eSNVs showed stronger performance than the logistic regression models for lincRNA eSNVs. However, using the non-linear Boosted Trees algorithm, the situation was reversed, and the Boosted Trees models for lincRNA eSNVs showed stronger performance than the models for protein-coding gene eSNVs. This suggests that the prediction of lincRNA somatic mutations may be more non-linear and more complex than the protein coding genes. The most important feature for the lincRNA somatic model was transversion (whether a mutation was a transversion – 1, or transition mutation – 0). Transition somatic mutations, particularly C>T transitions, are more frequent than transversion somatic mutations [37]. However, for particular tumor types and even specific genes (e.g., p53 somatic mutations),

the prevalence of transversions may be higher than transitions [37, 38]. This suggests the transversion or transition type of mutation may ultimately be important in determining a mutation's biological importance. In addition to the type of mutation, the fourth most important feature was CG_0 (i.e., whether the reference at the mutation site was a C/G nucleotide – 1 or a AT nucleotide – 0). These findings show that the information about the basic chemistry of the mutation is one of the strongest indication of mutation likelihood.

Conservation is the second most important feature in the lincRNA somatic model. Although conservation scores are determined through evolutionary homology, it has been shown that conservation correlates with somatic mutation hot spots [39]. The germline models for lincRNAs, in contrast, scored conservation as the most important feature. This may be expected, as conservation itself is a direct measure of the likelihood of variation through a species' germline lineage. The third most important feature for most lincRNA somatic models was cnv_pos (i.e., copy number variation at the mutation position, determined by a microarray on a corresponding DNA TCGA sample). Previous studies have found that many somatic gene mutations are significantly correlated with their mutation profiles and copy number alterations in cancer, including EGFR and KRAS [40]. However, although many genes were found to be correlated, on a global scale, many genes did not reach significance [40]. The cnv_pos feature was more important in the lincRNA models compared to the protein coding gene models, suggesting that cnv_pos may be a relatively more important feature for lincRNAs. Furthermore, cnv_pos was significantly less important in both germline models.

For the datasets with matched tissue cell line histone data, histone features related to the lincRNA sites were determined to have a significant effect on the prediction of eSNVs sites. Previous studies have found that chromatin modifications had a major effect on regional mutation rates in cancer cells [41]. Since histone methylation and acetylation status determines the 3-dimensional conformation and openness of genomic regions, differences in histone modifications between regions may change the exposure of a region to mutagenic forces.

Several additional features related to the coding frame of protein coding genes were included in the models predicting somatic protein coding gene eSNVs. Somatic mutations in the coding region of genes are much less frequent compared to non-coding portions, such as the UTR regions and introns [42]. In support of this idea, of these additional features added to the protein coding gene models, the most important was utr3 (whether a mutation came from a 3'UTR region). We showed that the mutation and expression profile of known lincRNA driver genes are statistically different than non-driver or unknown lincRNAs (Figure 6 and Figure S8). Despite higher expression being correlated with a higher chance of mutation (Figure S10), the lincRNA mutation likelihood is significantly lower for known driver lincRNAs, and the overall

expression for known driver lincRNAs is significantly higher. This corroborates previous studies that showed that known cancer driver mutations in non-coding regions actually have lower somatic and germline mutation scores [43].

While using RNA-Seq to perform mutation calling is an interesting idea to couple SNVs with expression data, false negatives may arise due to the fact that many lincRNAs and transcripts are lowly expressed or not expressed at all in certain tissues or conditions [44]. On the other hand, false positives may also be introduced as RNA splicing of transcripts could cause additional read misalignment to the genome reference [10]. Similarly, since expression data and eSNVs both come from RNA-Seq and require the presence of expressed transcripts to produce reads for measurement, expression and eSNVs are inherently coupled. A gene that is not expressed will also not have any detected mutations. This suggests that there may be bias towards regions of high read coverage and therefore high expression.

However, within the TCGA RNA-Seq datasets, the majority of eSNVs detected from RNA-Seq within exome probe bounadries, are also detected in exome-sequencing variant calling from the same patients (Figure S2 and Figure S3). Previous studies have found that, from the same patient, the concordance between sequencing platforms and variant calling software to be about 50% [27]. This suggests that the false positives from the eSNV RNA-Seq pipeline are much less of an issue than other technical factors, such as the choice of sequencing platform.

While the models built and used in this study have strong performance (AUC) on balanced datasets, the sparsity of SNPs and SNVs in a genome suggests that individual sites may not be able to be definitively predicted with high certainty. Biologically, this is a result of the stochastic nature of somatic point mutations. However, individual genes, lincRNAs, genomic regions, or possibly individual exons or sections of lincRNAs may be predicted as more or less likely to be mutated, relative to other exons or genes. This is an important step in finding the biological and tumorigenic significance of genes and lincRNAs that are susceptible or resistant to somatic mutations. In this study, we explored eSNV landscape in tumor samples, with a focus on non-coding regions (specifically lincRNAs).

At the individual SNV level, we showed that mutations can be accurately classified, and the probability of particular mutations can be estimated. At the lincRNA level, we showed that lincRNAs had a spectrum of mutation profiles and correlates with gene expression. Interestingly, lincRNAs that are known to be drivers in cancer tumorigenesis and function show markedly higher gene expression, yet lower mutation probabilities. Finally, we showed that the combined lincRNA mutation scores (average log odds of mutation) that integrates RNA-Seq, copy number variation, nucleotide composition, histone marker data etc., has more robust clustering than

just the expression data from RNA-Seq alone. In summary, we generate models capable of predicting mutation probabilities of individual lincRNA nucleotide positions as well as overall lincRNA mutations probabilities. This study advances the knowledge of the mutational forces of lincRNAs in comparison to protein coding genes.

References

1. Ching, T., Masaki, J., Weirather, J. & Garmire, L. X. Non-coding yet non-trivial: a review on the computational genomics of lincRNAs. *BioData mining* **8,** 1 (2015).

2. Ching, T., Peplowska, K., Huang, S., Zhu, X., Shen, Y., Molnar, J., Yu, H., Tiirikainen, M., Fogelgren, B. & Fan, R. Pan-Cancer Analyses Reveal Long Intergenic Non-Coding RNAs Relevant to Tumor Diagnosis, Subtyping and Prognosis. *EBioMedicine* (2016).

3. Prensner, J. R. & Chinnaiyan, A. M. The emergence of lncRNAs in cancer biology. *Cancer discovery* **1,** 391–407 (2011).

4. Yang, Y., Li, H., Hou, S., Hu, B., Liu, J. & Wang, J. The Noncoding RNA Expression Profile and the Effect of lncRNA AK126698 on Cisplatin Resistance in Non-Small-Cell Lung Cancer Cell. *PLOS ONE* **8,** e65309. ISSN: 1932-6203 (May 2013).

5. Zarate, R., Boni, V., Bandres, E. & Garcia-Foncillas, J. MiRNAs and LincRNAs: Could They Be Considered as Biomarkers in Colorectal Cancer? *International Journal of Molecular Sciences* **13,** 840–865 (Jan. 2012).

6. Zhou, X., Chen, J. & Tang, W. The molecular mechanism of HOTAIR in tumorigenesis, metastasis, and drug resistance. *Acta Biochimica et Biophysica Sinica* **46,** 1011–1015. ISSN: 1672-9145, 1745-7270 (Dec. 2014).

7. Gonzalez-Perez, A., Perez-Llamas, C., Deu-Pons, J., Tamborero, D., Schroeder, M. P., Jene-Sanz, A., Santos, A. & Lopez-Bigas, N. IntOGen-mutations identifies cancer drivers across tumor types. *Nature methods* **10,** 1081–1082 (2013).

8. Tamborero, D., Gonzalez-Perez, A., Perez-Llamas, C., Deu-Pons, J., Kandoth, C., Reimand, u., Lawrence, M. S., Getz, G., Bader, G. D., Ding, L. & et al. Comprehensive identification of mutational cancer driver genes across 12 tumor types. *Scientific Reports* **3.** ISSN: 2045-2322. doi:10.1038/srep02650 (Oct. 2013).

9. He, Q., He, Q., Liu, X., Wei, Y., Shen, S., Hu, X., Li, Q., Peng, X., Wang, L. & Yu, L. Genome-wide prediction of cancer driver genes based on SNP and cancer SNV data. *American journal of cancer research* **4,** 394 (2014).

10. Tang, X., Baheti, S., Shameer, K., Thompson, K. J., Wills, Q., Niu, N., Holcomb, I. N., Boutet, S. C., Ramakrishnan, R., Kachergus, J. M. & et al. The eSNV-detect: a computational system to identify expressed single nucleotide variants from transcriptome sequencing data. *Nucleic acids research,* gku1005 (2014).

11. Ng, S. B., Turner, E. H., Robertson, P. D., Flygare, S. D., Bigham, A. W., Lee, C., Shaffer, T., Wong, M., Bhattacharjee, A., Eichler, E. E. & et al. Targeted Capture and Massively Parallel Sequencing of Twelve Human Exomes. *Nature* **461,** 272–276. ISSN: 0028-0836 (Sept. 2009).

12. Warr, A., Robert, C., Hume, D., Archibald, A., Deeb, N. & Watson, M. Exome Sequencing: Current and Future Perspectives. *G3: Genes—Genomes—Genetics* **5,** 1543–1550. ISSN: , 2160-1836 (Aug. 2015).

13. Bryois, J., Buil, A., Evans, D. M., Kemp, J. P., Montgomery, S. B., Conrad, D. F., Ho, K. M., Ring, S., Hurles, M., Deloukas, P. & et al. Cis and trans effects of human genomic variants on gene expression. *PLoS Genet* **10,** e1004461 (2014).

14. Hu, P., Lan, H., Xu, W., Beyene, J. & Greenwood, C. M. Identifying cis-and trans-acting single-nucleotide polymorphisms controlling lymphocyte gene expression in humans. *BMC proceedings* **1,** 1 (2007).

15. Peng, Z., Cheng, Y., Tan, B. C.-M., Kang, L., Tian, Z., Zhu, Y., Zhang, W., Liang, Y., Hu, X., Tan, X. & et al. Comprehensive analysis of RNA-Seq data reveals extensive RNA editing in a human transcriptome. *Nature biotechnology* **30,** 253–260 (2012).

16. Piskol, R., Ramaswami, G. & Li, J. B. Reliable identification of genomic variants from RNA-seq data. *The American Journal of Human Genetics* **93,** 641–651 (2013).

17. Liao, Y., Smyth, G. K. & Shi, W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **30,** 923–930 (2014).

18. Consortium, E. P. *et al.* The ENCODE (ENCyclopedia of DNA elements) project. *Science* **306,** 636–640 (2004).

19. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome biology* **15,** 1 (2014).

20. Durinck, S., Spellman, P. T., Birney, E. & Huber, W. Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. *Nature protocols* **4,** 1184–1191 (2009).

21. Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M. & Gingeras, T. R. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29,** 15–21 (2013).

22. McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M. & et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome research* **20,** 1297–1303 (2010).

23. Chen, T. & Guestrin, C. Xgboost: A scalable tree boosting system. *Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining,* 785–794 (2016).

24. Dal Pozzolo, A., Caelen, O., Johnson, R. A. & Bontempi, G. Calibrating Probability with Undersampling for Unbalanced Classification. *Computational Intelligence, 2015 IEEE Symposium Series,* 159–166 (2015).

25. Strehl, A. & Ghosh, J. Cluster ensembles – a knowledge reuse framework for combining multiple partitions. *Journal of machine learning research* **3,** 583–617 (2002).

26. Cover, T. M. & Thomas, J. A. *Elements of information theory* (John Wiley & Sons, 2012).

27. O'Rawe, J., Jiang, T., Sun, G., Wu, Y., Wang, W., Hu, J., Bodily, P., Tian, L., Hakonarson, H., Johnson, W. E. & et al. Low concordance of multiple variant-calling pipelines: practical implications for exome and genome sequencing. *Genome medicine* **5,** 1 (2013).

28. Cornish, A. & Guda, C. A comparison of variant calling pipelines using genome in a bottle as a reference. *BioMed research international* **2015** (2015).

29. Sahin, A. A., Edgerton, M. E., Murray, J. L. & Bondy, M. Copy Number Imbalances between Screen-and Symptom-Detected Breast Cancers and Impact on Disease-Free Survival. *Cancer prevention research* (2011).

30. Ning, S., Zhang, J., Wang, P., Zhi, H., Wang, J., Liu, Y., Gao, Y., Guo, M., Yue, M., Wang, L. & et al. Lnc2Cancer: a manually curated database of experimentally supported lncRNAs associated with various human cancers. *Nucleic acids research,* gkv1094 (2015).

31. Ahn, Y.-Y., Bagrow, J. P. & Lehmann, S. Link communities reveal multiscale complexity in networks. *Nature* **466,** 761–764 (2010).

32. Smith, K. S., Yadav, V. K., Pei, S., Pollyea, D. A., Jordan, C. T. & De, S. SomVarIUS: somatic variant identification from unpaired tissue samples. *Bioinformatics,* btv685 (2015).

33. Meric-Bernstam, F., Johnson, A., Holla, V., Bailey, A. M., Brusco, L., Chen, K., Routbort, M., Patel, K. P., Zeng, J., Kopetz, S. & et al. A Decision Support Framework for Genomically Informed Investigational Cancer Therapy. *Journal of the National Cancer Institute* **107,** djv098. ISSN: 0027-8874, 1460-2105 (July 2015).

34. Costello, M., Pugh, T. J., Fennell, T. J., Stewart, C., Lichtenstein, L., Meldrim, J. C., Fostel, J. L., Friedrich, D. C., Perrin, D., Dionne, D. & et al. Discovery and characterization of artifactual mutations in deep coverage targeted capture sequencing data due to oxidative DNA damage during sample preparation. *Nucleic acids research,* gks1443 (2013).

35. Cibulskis, K., Lawrence, M. S., Carter, S. L., Sivachenko, A., Jaffe, D., Sougnez, C., Gabriel, S., Meyerson, M., Lander, E. S. & Getz, G. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nature biotechnology* **31,** 213–219 (2013).

36. Heppner, G. H., Dexter, D. L., DeNucci, T., Miller, F. R. & Calabresi, P. Heterogeneity in drug sensitivity among tumor cell subpopulations of a single mammary tumor. *Cancer research* **38,** 3758–3763 (1978).

37. Kandoth, C., McLellan, M. D., Vandin, F., Ye, K., Niu, B., Lu, C., Xie, M., Zhang, Q., McMichael, J. F., Wyczalkowski, M. A. & et al. Mutational landscape and significance across 12 major cancer types. *Nature* **502,** 333–339 (2013).

38. Hollstein, M., Sidransky, D., Vogelstein, B. & Harris, C. C. p53 mutations in human cancers. *Science* **253,** 49–54 (1991).

39. Walker, R., Bond, J. P., Tarone, R. E., Harris, C. C., Makalowski, W., Boguski, M. S. & Greenblatt, M. S. Evolutionary conservation and somatic mutation hotspot maps of p53: correlation with p53 protein structural and functional features. *Oncogene* **18,** 211–218 (1999).

40. Ding, L., Getz, G., Wheeler, D. A., Mardis, E. R., McLellan, M. D., Cibulskis, K., Sougnez, C., Greulich, H., Muzny, D. M., Morgan, M. B. & et al. Somatic mutations affect key pathways in lung adenocarcinoma. *Nature* **455,** 1069–1075 (2008).

41. Schuster-Bockler, B. & Lehner, B. Chromatin organization is a major influence on regional mutation rates in human cancer cells. *nature* **488,** 504–507 (2012).

42. Lee, W., Jiang, Z., Liu, J., Haverty, P. M., Guan, Y., Stinson, J., Yue, P., Zhang, Y., Pant, K. P., Bhatt, D., *et al.* The mutation spectrum revealed by paired genome sequences from a lung cancer patient. *Nature* **465,** 473–477 (2010).

43. Li, J., Poursat, M.-A., Drubay, D., Motz, A., Saci, Z., Morillon, A., Michiels, S. & Gautheret, D. A dual model for prioritizing cancer mutations in the non-coding genome based on germline and somatic events. *PLoS Comput Biol* **11,** e1004583 (2015).

44. Cabili, M. N., Trapnell, C., Goff, L., Koziol, M., Tazon-Vega, B., Regev, A. & Rinn, J. L. Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. *Genes & development* **25,** 1915–1927 (2011).

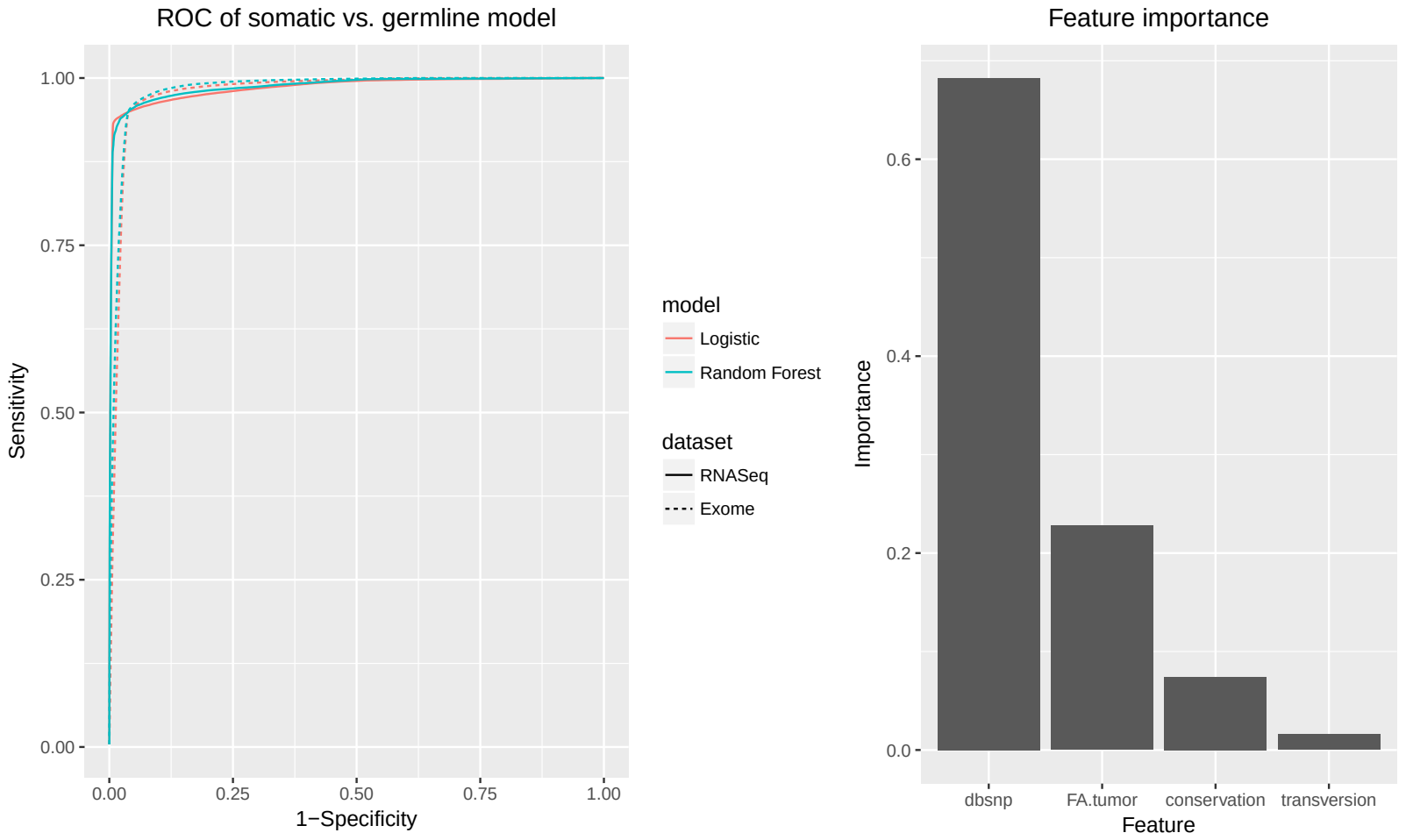


Figure 1. A – Receiver operating characteristic (ROC) curve showing the performance of the random forest and logistic regression models for differentiating somatic and germline mutations in the exome-sequencing data. B – Feature importance based on the random forest model.

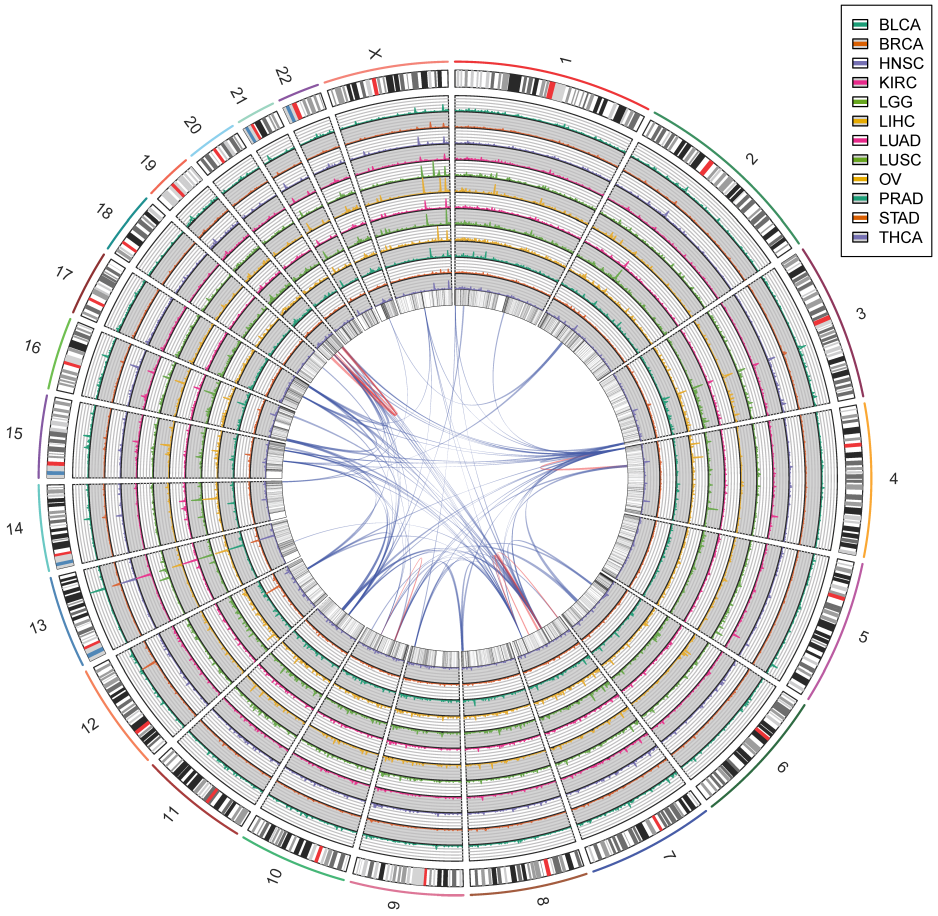


Figure 2. Circle plot of the normalized lincRNA eSNV mutation density. For each cancer type, the number of lincRNA eSNVs was binned using window sizes of 100,000 across the genome. The bin count was then normalized by the lincRNA exon density in the corresponding transcriptomic region. The outer layer shows the human genome cytogenetics. The inner layers are the lincRNA eSNV densities from the datasets in order (from outermost to innermost): BLCA, BRCA, HNSC, KIRC, LGG, LIHC, LUAD, LUSC, OV, PRAD, STAD and THCA. The inner layer shows the exon density of the lincRNA transcriptome.

Figure 3. Model performance differentiating somatic eSNVs from background in lincRNAs and protein coding genes for the 12 TCGA datasets, using the Gradient boosted Trees machine learning algorithm.

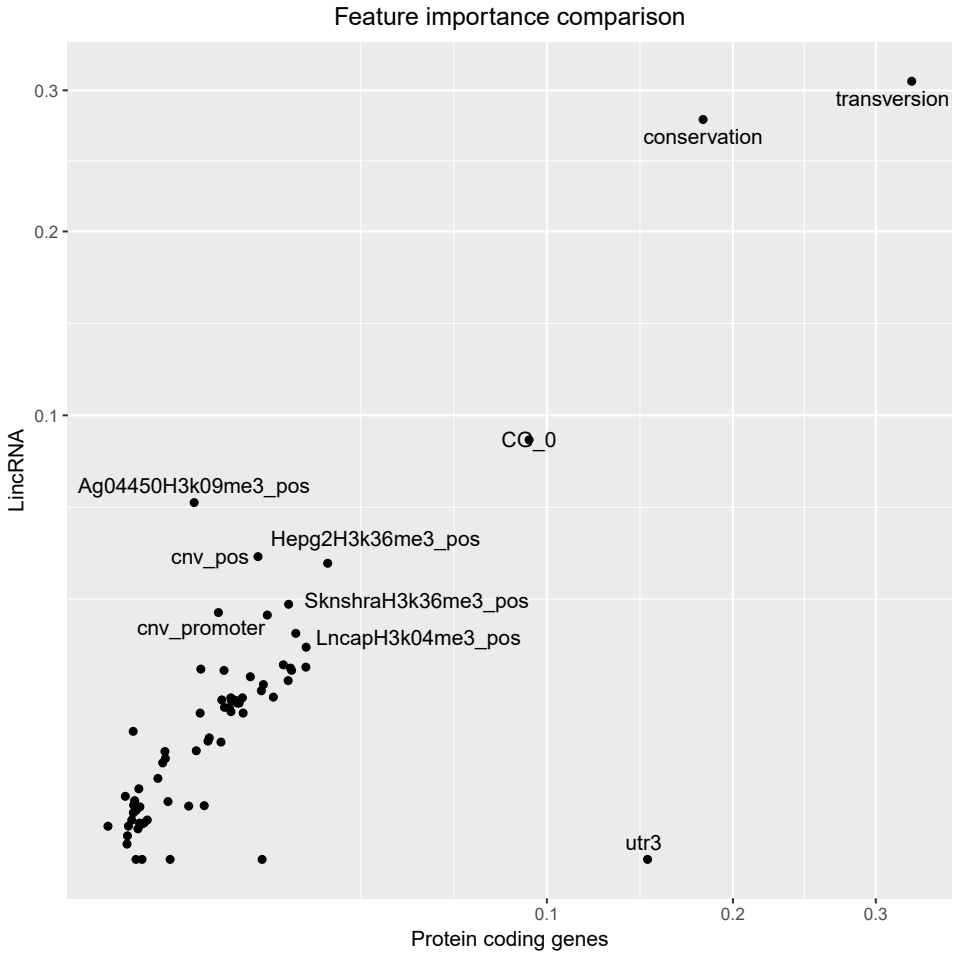


Figure 4. Feature importance for lincRNA and protein coding gene eSNVs. Feature importance for somatic eSNV models. Feature importance was calculated using the Gain measure, which evaluates the average increase in accuracy at a feature's node splitting in each tree.

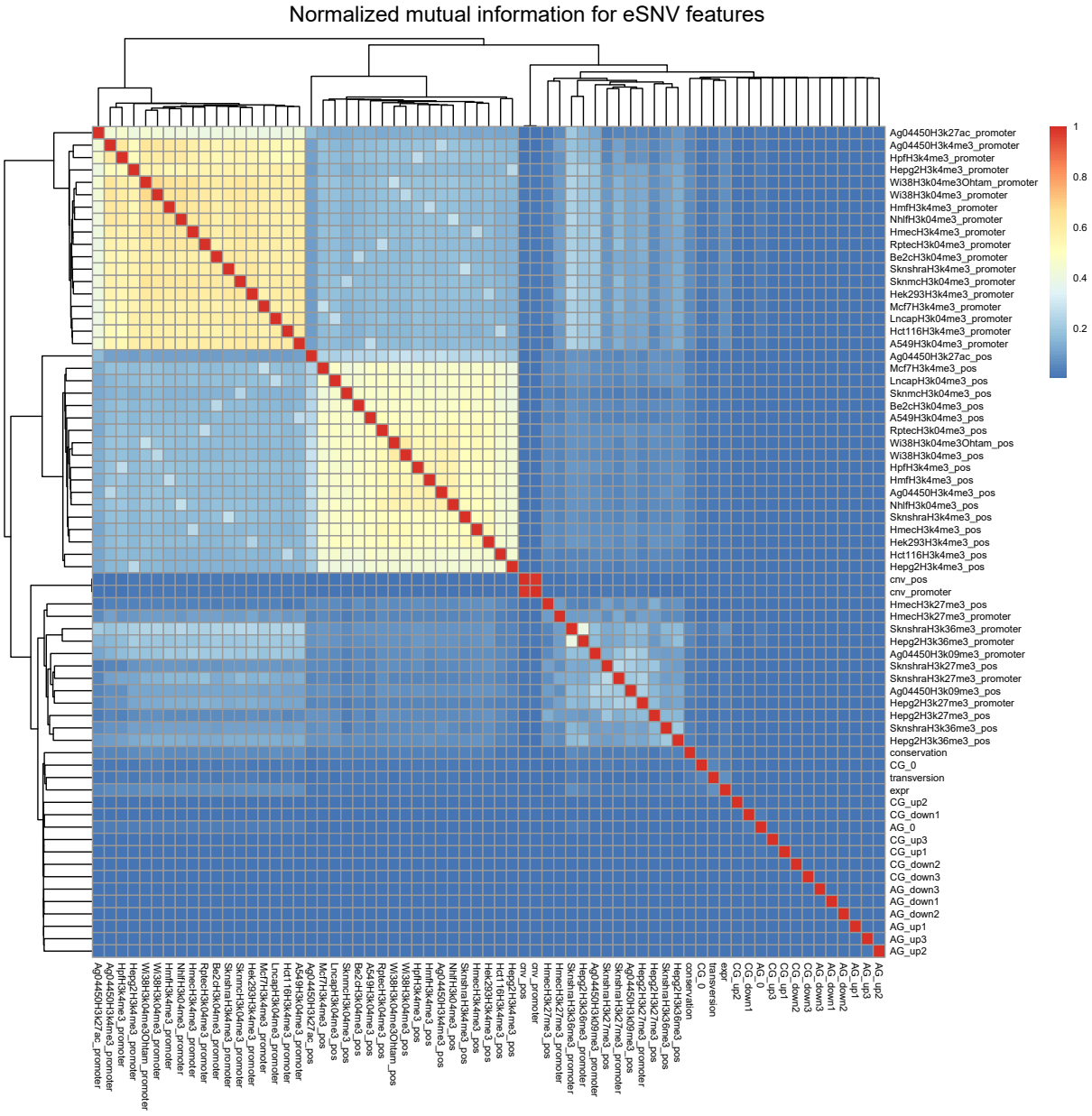


Figure 5. Heatmap of normalized mutual information between features in the eSNV lincRNA model.

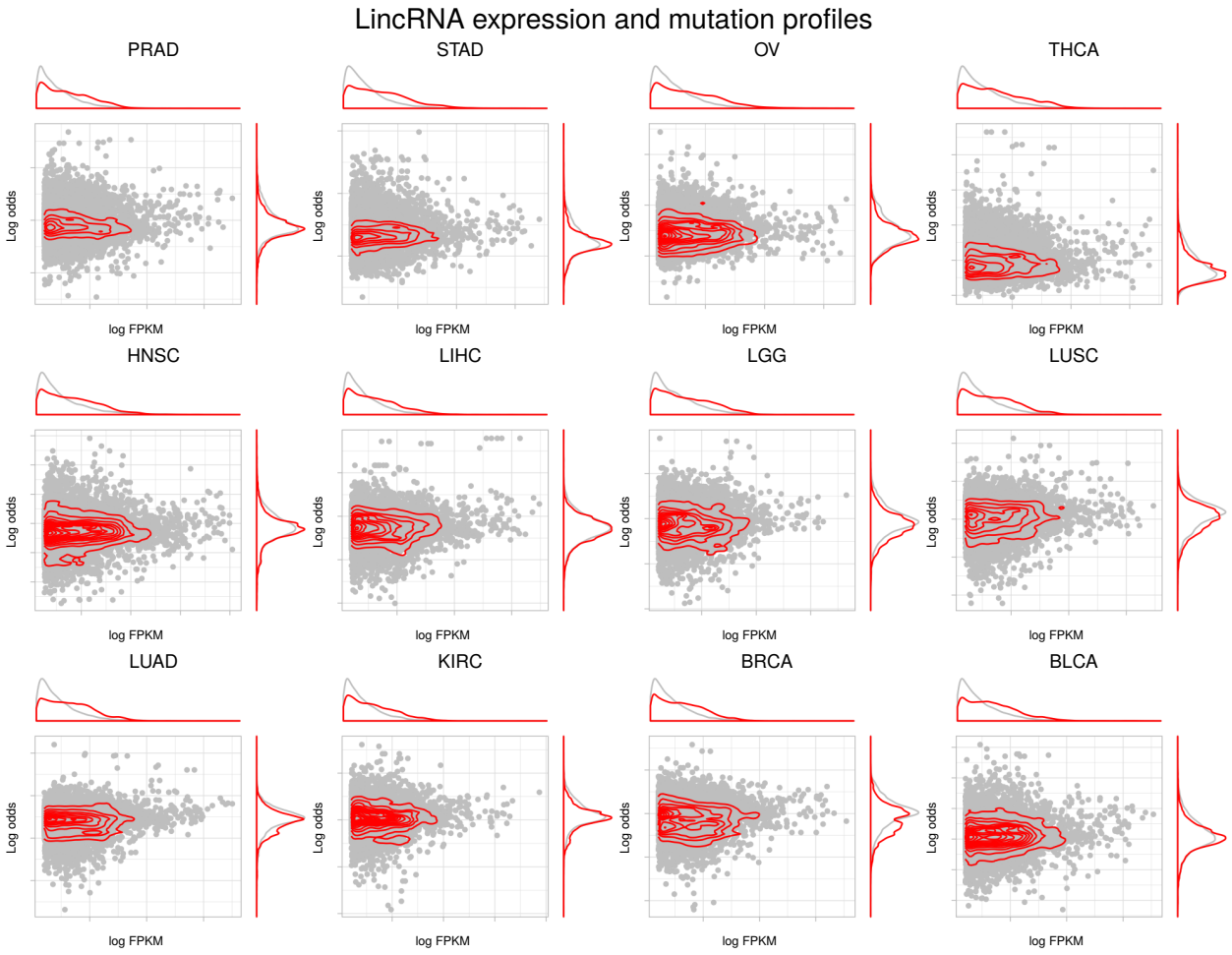


Figure 6. Summarized mutation score and expression profiles for known lincRNA drivers (red) and non-driver/unknown lincRNAs (gray) in each of the 12 TCGA datasets. Mutation scores were summarized as the mean log odds ratio over all base pairs of a lincRNA. Expression was calculated as the log of the fragments per kilobase per million reads (FPKM) for each lincRNA.

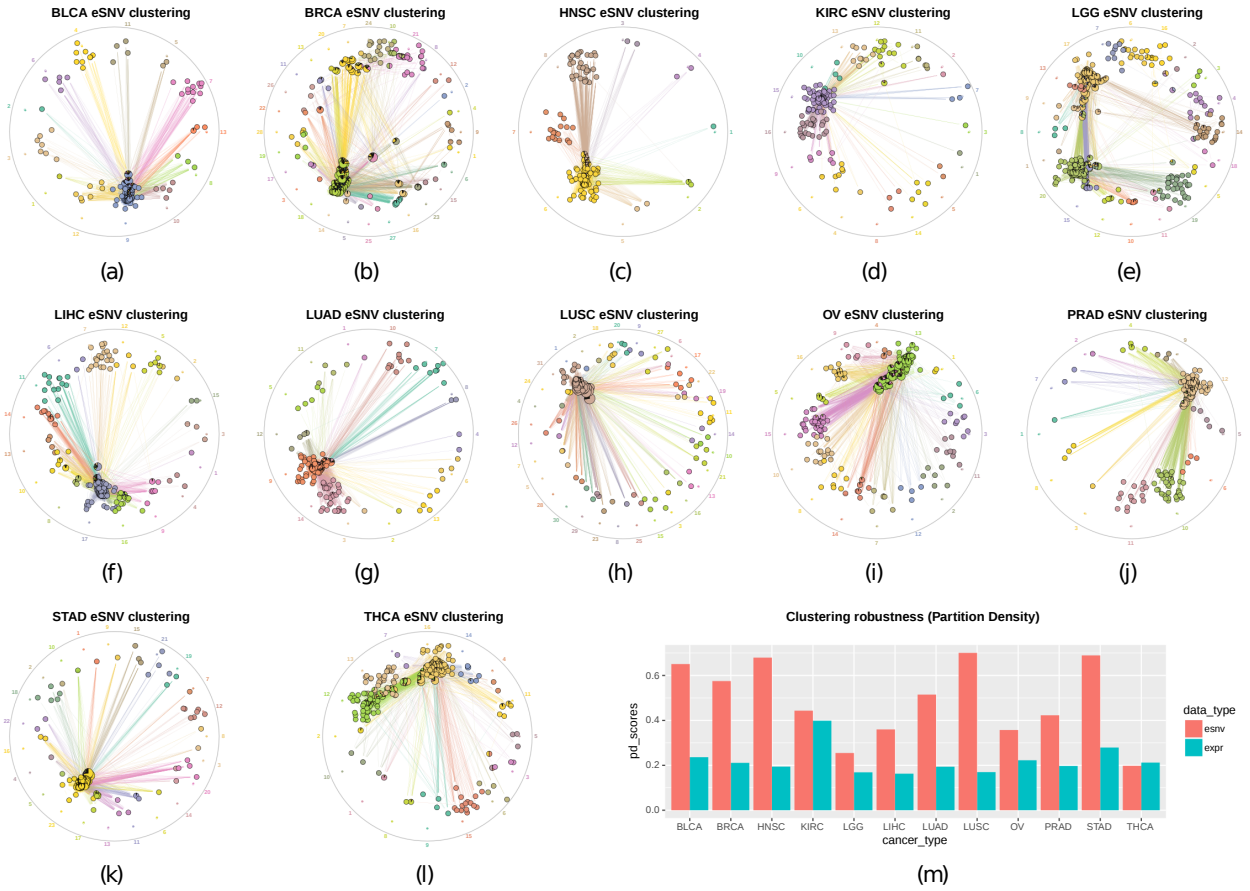


Figure 7. A-L – Clustering results for the 12 TCGA datasets, based on the lincRNA summarized eSNV score (mean log odds ratio of somatic mutation). Clustering was performed using the link community clustering method. M-partition density measure for comparing the clustering robustness between eSNV data and expression data.

6.6 Appendix

6.6.1 Supplemental figures and tables

Table S1. Patient and eSNV and germline variant tabulations for the 10 TCGA datasets.

Table S2. eSNV and germline variant density based on mutation calls from RNA-Seq data.

Table S3. Description of features of the datasets and models used in this study.

Figure S1. Flow chart diagram describing the pipeline for the raw data processing, calling eSNVs and calculating expression data.

Figure S2. Comparison barplots between RNA-Seq and exome-seq data from the same patients.

Figure S3. Heatmaps of the concordant variants detected in the RNA-Seq and exome-seq data.

Figure S4. Model performance differentiating germline variants or eSNVs from background in the 12 TCGA datasets.

Figure S5. Feature importance for germline variant models for lincRNAs and protein-coding genes.

Figure S6. Base pair distribution for detected eSNV models.

Figure S7. Feature distribution histograms for eSNV models.

Figure S8. Average distribution of mutation scores and expression for known lincRNA drivers and non-driver/unknown lincRNAs. Both expression and mutation scores were significantly different ($p < 2.2e\text{-}16$).

Figure S9. Link community clustering results for the 12 TCGA datasets, based on the lincRNA expression data.

Figure S10. LincRNA expression distribution for detected eSNVs and negative background.

Supplemental methods. Supplemental information regarding the processing and filtering of the eSNVs for variant calling.

Table S1

cancer_type	Patients	Total variants	Protein-coding eSNV	LincRNA eSNV	Protein-coding germline	LincRNA germline
PRAD	491	33240.43 +/- 5951.55	693.98 +/- 235.50	106.92 +/- 37.44	24996.55 +/- 4292.20	2292.13 +/- 580.21
STAD	414	41056.52 +/- 7581.96	1178.52 +/- 483.83	151.19 +/- 82.21	29852.28 +/- 4825.27	3260.55 +/- 1218.17
OV	300	42750.73 +/- 6373.53	1207.35 +/- 310.14	225.18 +/- 106.01	30285.15 +/- 3770.79	3992.69 +/- 1113.04
THCA	497	34568.26 +/- 5288.24	760.66 +/- 171.73	116.07 +/- 38.49	25974.94 +/- 3788.85	2381.96 +/- 592.03
HNSC	514	29939.30 +/- 5093.94	641.67 +/- 188.29	59.34 +/- 25.45	22786.42 +/- 3672.31	1635.16 +/- 477.41
LIHC	364	25213.55 +/- 5372.64	548.04 +/- 151.57	62.30 +/- 23.87	19119.57 +/- 3981.40	1512.20 +/- 469.33
LGG	513	37286.35 +/- 4488.98	834.35 +/- 366.00	137.00 +/- 37.55	27651.25 +/- 3098.94	2897.41 +/- 526.35
LUSC	498	33415.90 +/- 4734.57	798.53 +/- 230.31	91.63 +/- 39.80	24948.94 +/- 3140.12	2243.06 +/- 608.68
LUAD	512	33001.78 +/- 5801.08	746.25 +/- 259.69	83.62 +/- 44.04	24595.81 +/- 3847.28	2084.84 +/- 709.07
KIRC	525	36731.71 +/- 6346.44	847.07 +/- 248.17	120.83 +/- 52.79	27061.65 +/- 4322.88	2710.20 +/- 780.21
BRCA	1084	34589.13 +/- 5619.00	875.70 +/- 286.37	108.86 +/- 46.07	25571.13 +/- 3813.70	2251.74 +/- 621.48
BLCA	406	29838.92 +/- 5182.94	742.48 +/- 242.29	78.19 +/- 33.93	22210.54 +/- 3614.80	1909.15 +/- 584.05
All	6118	34235.03 +/- 6927.45	817.49 +/- 320.50	108.77 +/- 61.39	25407.92 +/- 4663.37	2379.47 +/- 903.91

Table S2

cancer_type	Overall variants	Protein-coding eSNV	LincRNA eSNV	Protein-coding germline	LincRNA germline
PRAD	2.63e-04 +/- 4.71e-05	7.85e-06 +/- 2.66e-06	2.82e-06 +/- 9.88e-07	2.83e-04 +/- 4.86e-05	6.05e-05 +/- 1.53e-05
STAD	3.25e-04 +/- 6.00e-05	1.33e-05 +/- 5.48e-06	3.99e-06 +/- 2.17e-06	3.38e-04 +/- 5.46e-05	8.60e-05 +/- 3.21e-05
OV	3.39e-04 +/- 5.05e-05	1.37e-05 +/- 3.51e-06	5.94e-06 +/- 2.80e-06	3.43e-04 +/- 4.27e-05	1.05e-04 +/- 2.94e-05
THCA	2.74e-04 +/- 4.19e-05	8.61e-06 +/- 1.94e-06	3.06e-06 +/- 1.02e-06	2.94e-04 +/- 4.29e-05	6.29e-05 +/- 1.56e-05
HNSC	2.37e-04 +/- 4.03e-05	7.26e-06 +/- 2.13e-06	1.57e-06 +/- 6.72e-07	2.58e-04 +/- 4.16e-05	4.31e-05 +/- 1.26e-05
LIHC	2.00e-04 +/- 4.25e-05	6.20e-06 +/- 1.72e-06	1.64e-06 +/- 6.30e-07	2.16e-04 +/- 4.51e-05	3.99e-05 +/- 1.24e-05
LGG	2.95e-04 +/- 3.56e-05	9.44e-06 +/- 4.14e-06	3.62e-06 +/- 9.91e-07	3.13e-04 +/- 3.51e-05	7.65e-05 +/- 1.39e-05
LUSC	2.65e-04 +/- 3.75e-05	9.04e-06 +/- 2.61e-06	2.42e-06 +/- 1.05e-06	2.82e-04 +/- 3.55e-05	5.92e-05 +/- 1.61e-05
LUAD	2.61e-04 +/- 4.59e-05	8.44e-06 +/- 2.94e-06	2.21e-06 +/- 1.16e-06	2.78e-04 +/- 4.35e-05	5.50e-05 +/- 1.87e-05
KIRC	2.91e-04 +/- 5.03e-05	9.59e-06 +/- 2.81e-06	3.19e-06 +/- 1.39e-06	3.06e-04 +/- 4.89e-05	7.15e-05 +/- 2.06e-05
BRCA	2.74e-04 +/- 4.45e-05	9.91e-06 +/- 3.24e-06	2.87e-06 +/- 1.22e-06	2.89e-04 +/- 4.32e-05	5.94e-05 +/- 1.64e-05
BLCA	2.36e-04 +/- 4.10e-05	8.40e-06 +/- 2.74e-06	2.06e-06 +/- 8.95e-07	2.51e-04 +/- 4.09e-05	5.04e-05 +/- 1.54e-05
All	2.71e-04 +/- 5.49e-05	9.25e-06 +/- 3.63e-06	2.87e-06 +/- 1.62e-06	2.88e-04 +/- 5.28e-05	6.28e-05 +/- 2.39e-05

*Units are variants per bp

Table S3 part 1

SNV position features

Feature	Description
conservation	PhyloP conservation score downloaded from UCSC genome browser
transversion	The variant is a transversion mutation - 1, or a transition mutation - 0
cnv_pos	Copy Number Variation from the matched DNA tumor sample in TCGA at the SNV position
CG_0	C or G nucleotide at the SNV position - 1, otherwise - 0
AG_0	A or G nucleotide at the SNV position - 1, otherwise - 0
CG_down1	C or G nucleotide 1 bp downstream - 1, otherwise - 0
AG_down1	A or G nucleotide 1 bp downstream - 1, otherwise - 0
CG_down2	C or G nucleotide 2 bp downstream - 1, otherwise - 0
AG_down2	A or G nucleotide 2 bp downstream - 1, otherwise - 0
CG_down3	C or G nucleotide 3 bp downstream - 1, otherwise - 0
AG_down3	A or G nucleotide 3 bp downstream - 1, otherwise - 0
CG_up1	C or G nucleotide 1 bp upstream - 1, otherwise - 0
AG_up1	A or G nucleotide 1 bp upstream - 1, otherwise - 0
CG_up2	C or G nucleotide 2 bp upstream - 1, otherwise - 0
AG_up2	A or G nucleotide 2 bp upstream - 1, otherwise - 0
CG_up3	C or G nucleotide 3 bp upstream - 1, otherwise - 0
AG_up3	A or G nucleotide 3 bp upstream - 1, otherwise - 0
A549H3k04me3_pos	H3K4 trimethylation at the SNV position of the A549 cell line, used in lung cancer datasets
Ag04450H3k09me3_pos	H3K9 trimethylation at the SNV position of the AG04450 cell line, used in lung cancer datasets
Ag04450H3k27ac_pos	H3K27 acetylation at the SNV position of the AG04450 cell line, used in lung cancer datasets
Ag04450H3k4me3_pos	H3K4 trimethylation at the SNV position of the AG04450 cell line, used in lung cancer datasets
Be2cH3k04me3_pos	H3K4 trimethylation at the SNV position of the BE2C cell line, used in the LGG dataset
Hek293H3k4me3_pos	H3K4 trimethylation at the SNV position of the HEK293 cell line, used in the KIRC dataset
Hepg2H3k27me3_pos	H3K27 trimethylation at the SNV position of the HEPG2 cell line, used in the LIHC dataset
Hepg2H3k36me3_pos	H3K36 trimethylation at the SNV position of the HEPG2 cell line, used in the LIHC dataset
Hepg2H3k4me3_pos	H3K4 trimethylation at the SNV position of the HEPG2 cell line, used in the LIHC dataset
HmecH3k27me3_pos	H3K27 trimethylation at the SNV position of the HMEC cell line, used in the BRCA dataset
HmecH3k4me3_pos	H3K4 trimethylation at the SNV position of the HMEC cell line, used in the BRCA dataset
HmfH3k4me3_pos	H3K4 trimethylation at the SNV position of the HMF cell line, used in the BRCA dataset
HpfH3k4me3_pos	H3K4 trimethylation at the SNV position of the HPF cell line, used in the lung cancer datasets

Table S3 part 2

Mcf7H3k4me3_pos	H3K4 trimethylation at the SNV position of the MCF7 cell line, used in the BRCA dataset
NhlfH3k04me3_pos	H3K4 trimethylation at the SNV position of the NHLF cell line, used in the lung cancer datasets
RptecH3k04me3_pos	H3K4 trimethylation at the SNV position of the RPTEC cell line, used in the KIRC dataset
SknmcH3k04me3_pos	H3K4 trimethylation at the SNV position of the SK-N-MC cell line, used in the LGG dataset
SknshraH3k27me3_pos	H3K27 trimethylation at the SNV position of the SK-N-SH_RA cell line, used in the LGG dataset
SknshraH3k36me3_pos	H3K36 trimethylation at the SNV position of the SK-N-SH_RA cell line, used in the LGG dataset
SknshraH3k4me3_pos	H3K4 trimethylation at the SNV position of the SK-N-SH_RA cell line, used in the LGG dataset
Wi38H3k04me3_pos	H3K4 trimethylation at the SNV position of the WI-38 cell line, used in the lung cancer datasets
Wi38H3k04me3Ohtam_pos	H3K4 trimethylation at the SNV position of the WI-38_Ohtam cell line, used in the lung cancer datasets

Promoter and gene-level features

Feature	Description
cnv_promoter	Copy Number Variation from the matched DNA tumor sample in TCGA at the promoter (TSS1500)
A549H3k04me3_promoter	H3K4 trimethylation at the promoter region of the A549 cell line, used in lung cancer datasets
Ag04450H3k09me3_promoter	H3K9 trimethylation at the promoter region of the AG04450 cell line, used in lung cancer datasets
Ag04450H3k27ac_promoter	H3K27 acetylation at the promoter region of the AG04450 cell line, used in lung cancer datasets
Ag04450H3k4me3_promoter	H3K4 trimethylation at the promoter region of the AG04450 cell line, used in lung cancer datasets
Be2cH3k04me3_promoter	H3K4 trimethylation at the promoter region of the BE2C cell line, used in the LGG dataset
Hek293H3k4me3_promoter	H3K4 trimethylation at the promoter region of the HEK293 cell line, used in the KIRC dataset
Hepg2H3k27me3_promoter	H3K27 trimethylation at the promoter region of the HEPG2 cell line, used in the LIHC dataset
Hepg2H3k36me3_promoter	H3K36 trimethylation at the promoter region of the HEPG2 cell line, used in the LIHC dataset
Hepg2H3k4me3_promoter	H3K4 trimethylation at the promoter region of the HEPG2 cell line, used in the LIHC dataset
HmecH3k27me3_promoter	H3K27 trimethylation at the promoter region of the HMEC cell line, used in the BRCA dataset
HmecH3k4me3_promoter	H3K4 trimethylation at the promoter region of the HMEC cell line, used in the BRCA dataset
HmfH3k4me3_promoter	H3K4 trimethylation at the promoter region of the HMF cell line, used in the BRCA dataset

Table S3 part 3

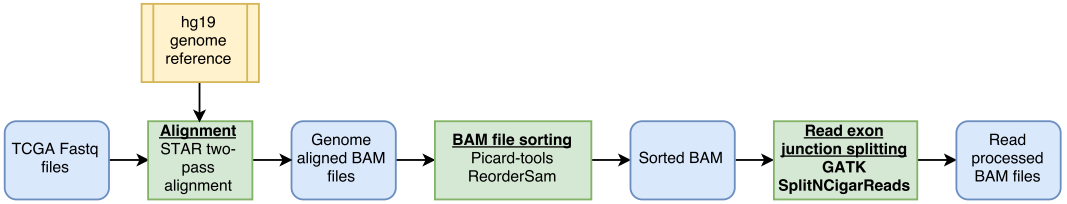
HpfH3k4me3_promoter	H3K4 trimethylation at the promoter region of the HPF cell line, used in the lung cancer datasets
Mcf7H3k4me3_promoter	H3K4 trimethylation at the promoter region of the MCF7 cell line, used in the BRCA dataset
NhlfH3k04me3_promoter	H3K4 trimethylation at the promoter region of the NHLF cell line, used in the lung cancer datasets
RptecH3k04me3_promoter	H3K4 trimethylation at the promoter region of the RPTEC cell line, used in the KIRC dataset
SknmcH3k04me3_promoter	H3K4 trimethylation at the promoter region of the SK-N-MC cell line, used in the LGG dataset
SknshraH3k27me3_promoter	H3K27 trimethylation at the promoter region of the SK-N-SH_RA cell line, used in the LGG dataset
SknshraH3k36me3_promoter	H3K36 trimethylation at the promoter region of the SK-N-SH_RA cell line, used in the LGG dataset
SknshraH3k4me3_promoter	H3K4 trimethylation at the promoter region of the SK-N-SH_RA cell line, used in the LGG dataset
Wi38H3k04me3_promoter	H3K4 trimethylation at the promoter region of the WI-38 cell line, used in the lung cancer datasets
Wi38H3k04me3Ohtam_promoter	H3K4 trimethylation at the promoter region of the WI-38_Ohtam cell line, used in the lung cancer datasets

Protein coding genes SNV additional features

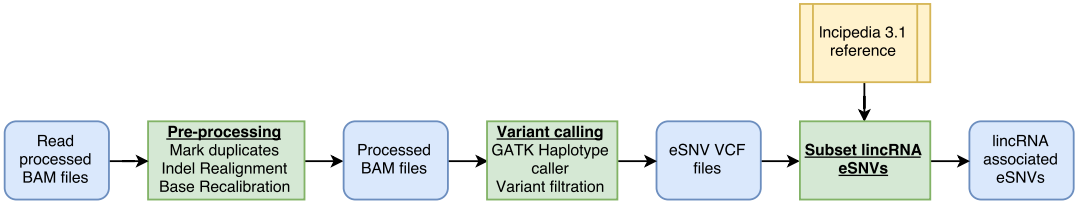
Feature	Description
nonsyn	Non-synonymous coding mutation - 1, other - 0
nonsense	Nonsense mutation - 1, other - 0
nonstop	Nonstop mutation - 1, other - 0
utr5	The SNV occurs in the 5'UTR region of the protein coding gene - 1, otherwise - 0
utr3	The SNV occurs in the 3'UTR region of the protein coding gene - 1, otherwise - 0

Figure S1

(1) RNA-Seq alignment and read pre-processing



(2) Re-alignment, re-calibration and lincRNA variant calling



(3) LincRNA expression quantification

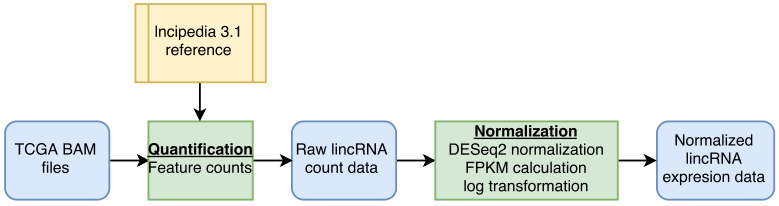


Figure S2

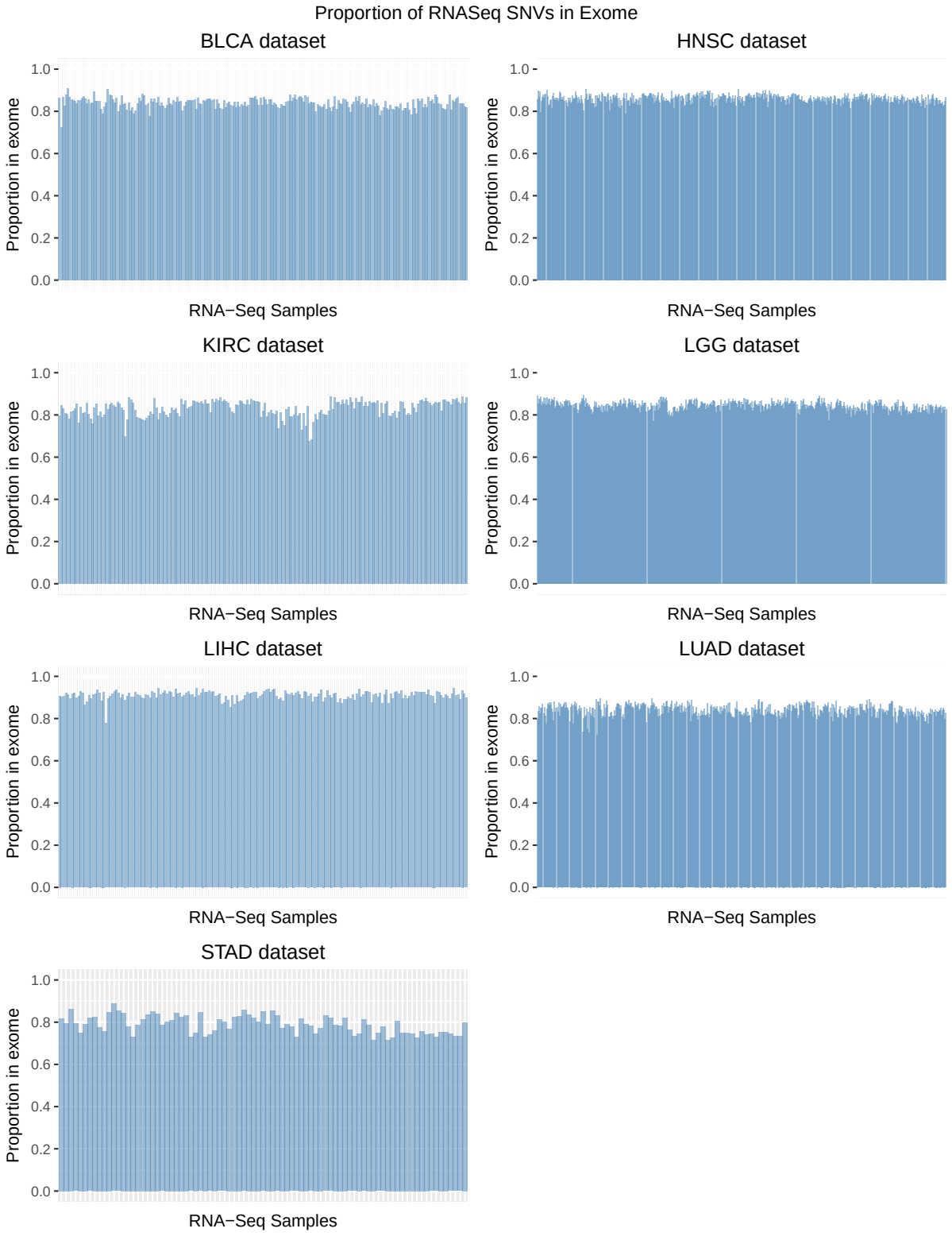
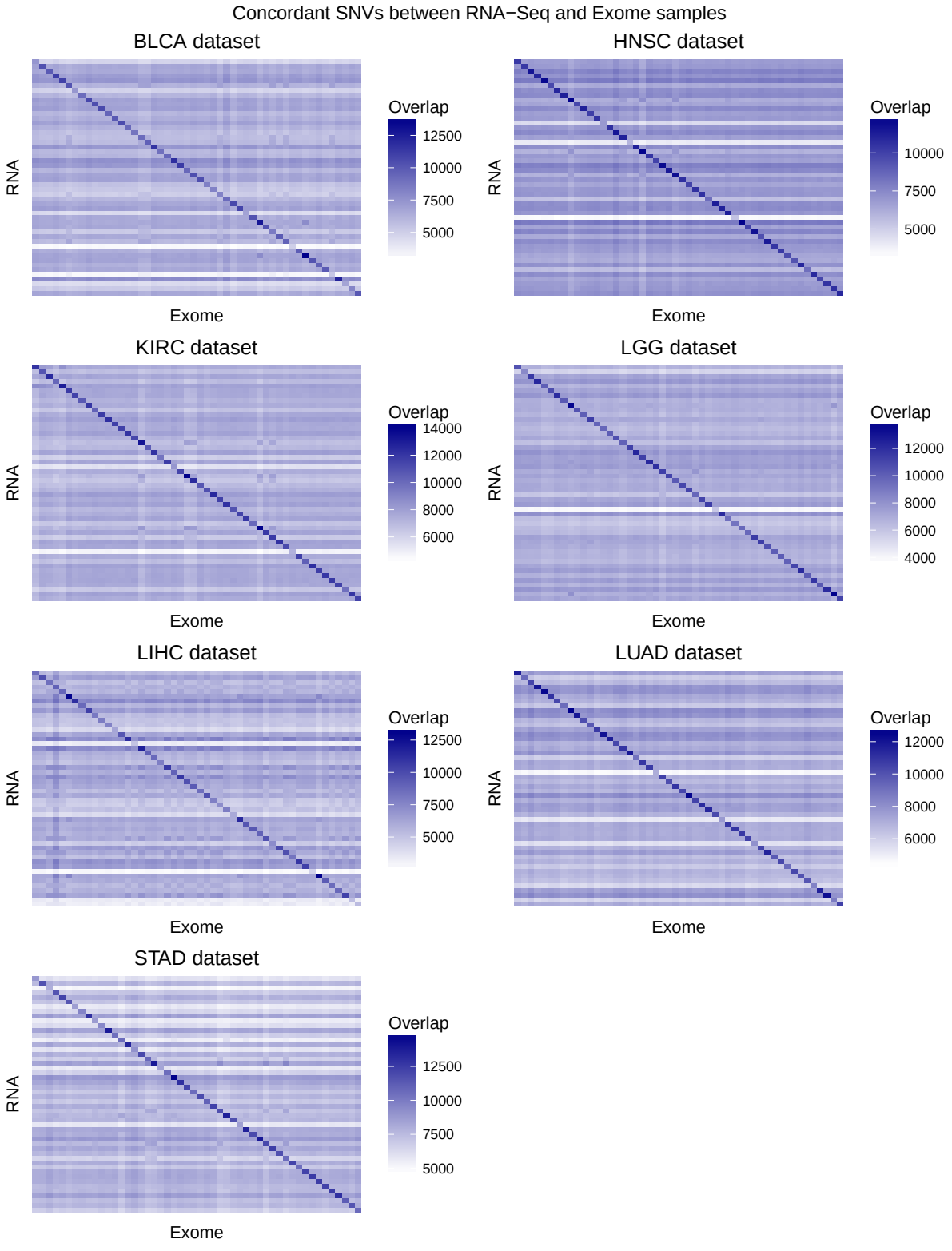

Proportion of RNASeq SNVs in Exome
BLCA dataset
HNSC dataset
KIRC dataset
LGG dataset
LIHC dataset
LUAD dataset
STAD dataset
Proportion in exome
RNA−Seq Samples
1.0
0.8
0.6
0.4
0.2
0.0

Figure S3


Concordant SNVs between RNA−Seq and Exome samples
BLCA dataset
HNSC dataset
KIRC dataset
LGG dataset
LIHC dataset
LUAD dataset
STAD dataset
RNA
Exome
Overlap
12500
10000
7500
5000
14000
12000
8000
6000
4000

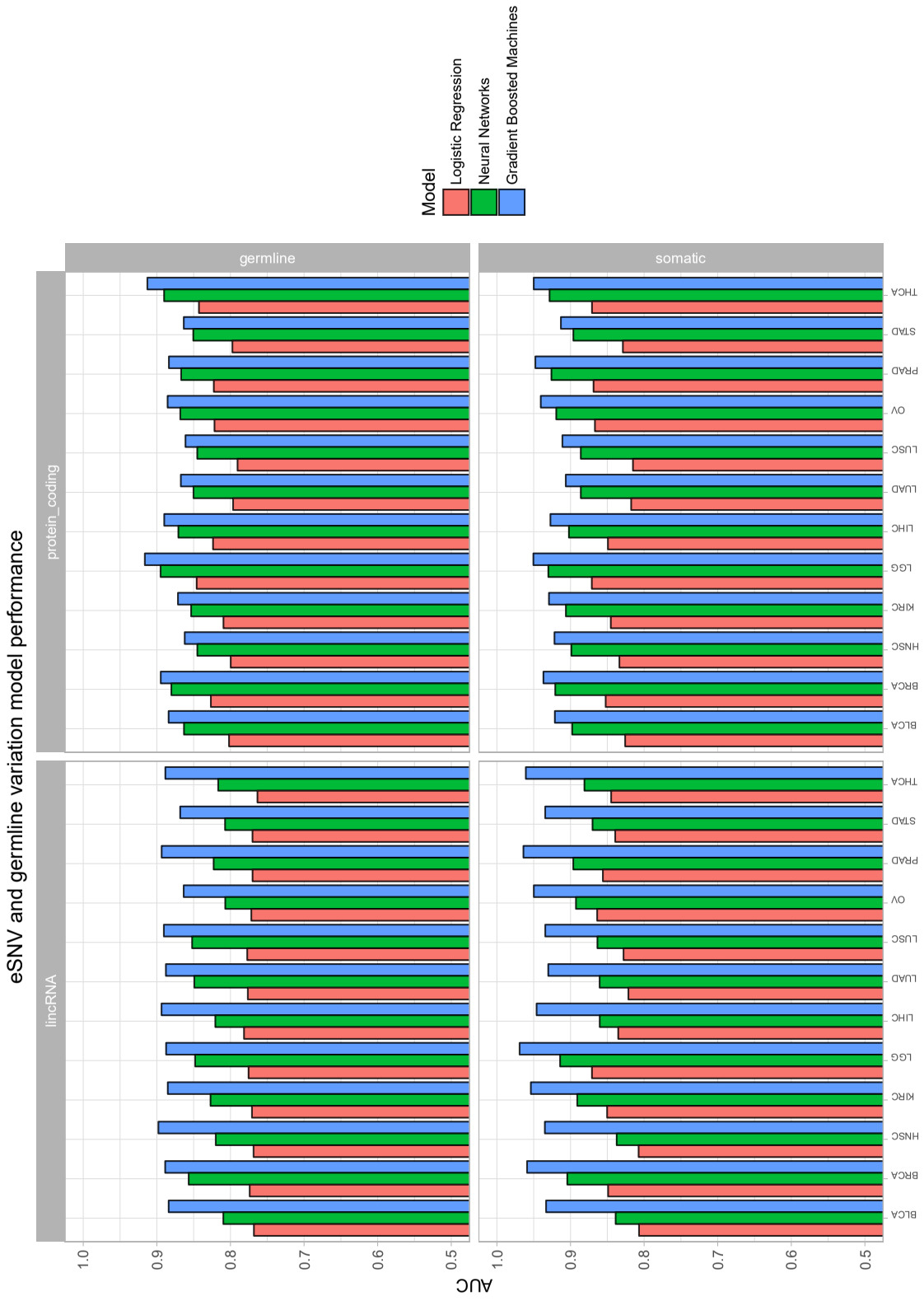

eSNV and germline variation model performance
Model
Logistic Regression
Neural Networks
Gradient Boosted Machines
germline
somatic
protein_coding
lincRNA
AUC
1.0
0.9
0.8
0.7
0.6
0.5
BLCA
BRCA
HNSC
KIRC
LGG
LIHC
LUAD
LUSC
OV
PRAD
STAD
THCA

Figure S4

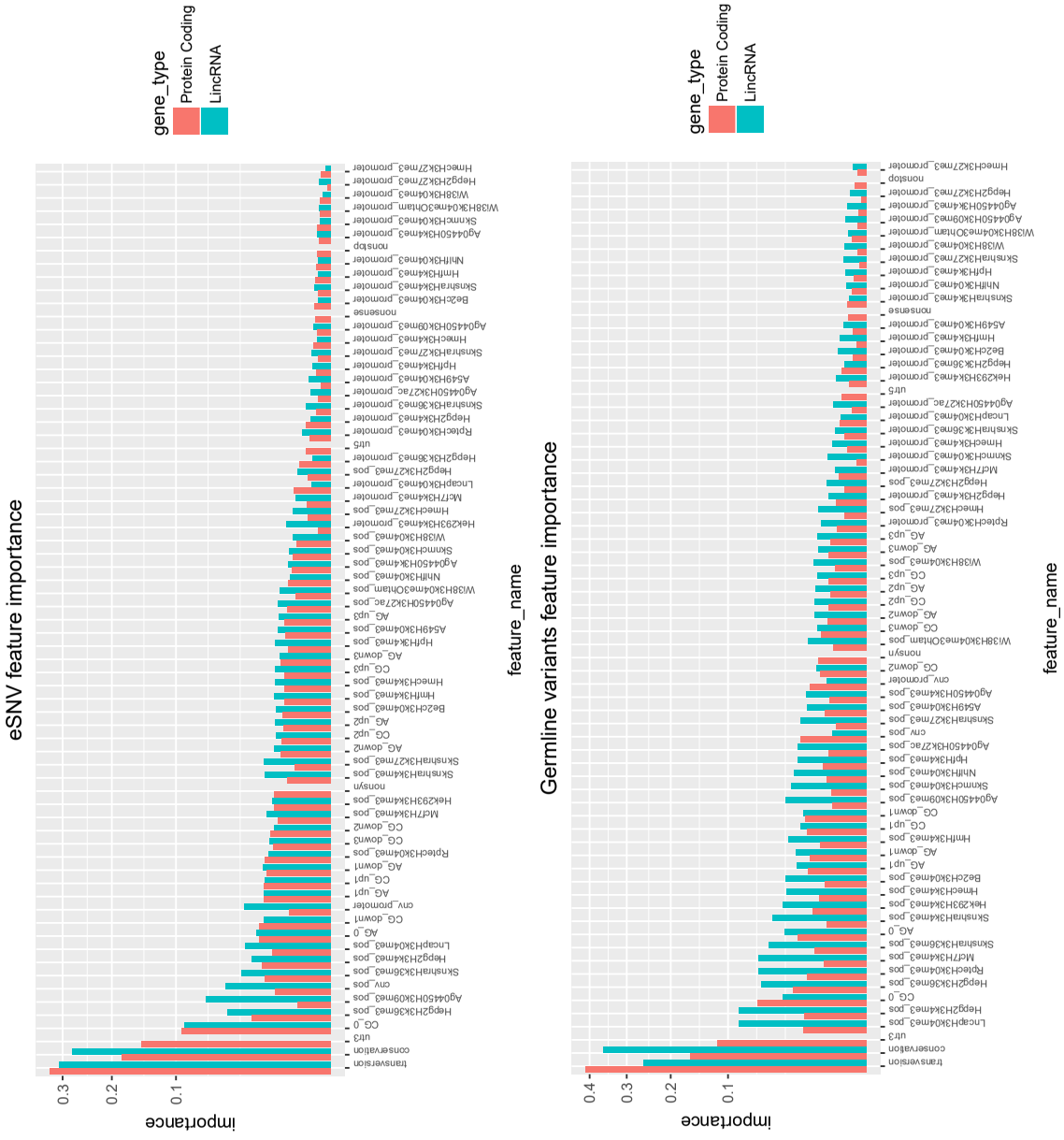

eSNV feature importance
Germline variants feature importance
gene_type
Protein Coding
LincRNA
importance
feature_name

Figure S5

Figure S6

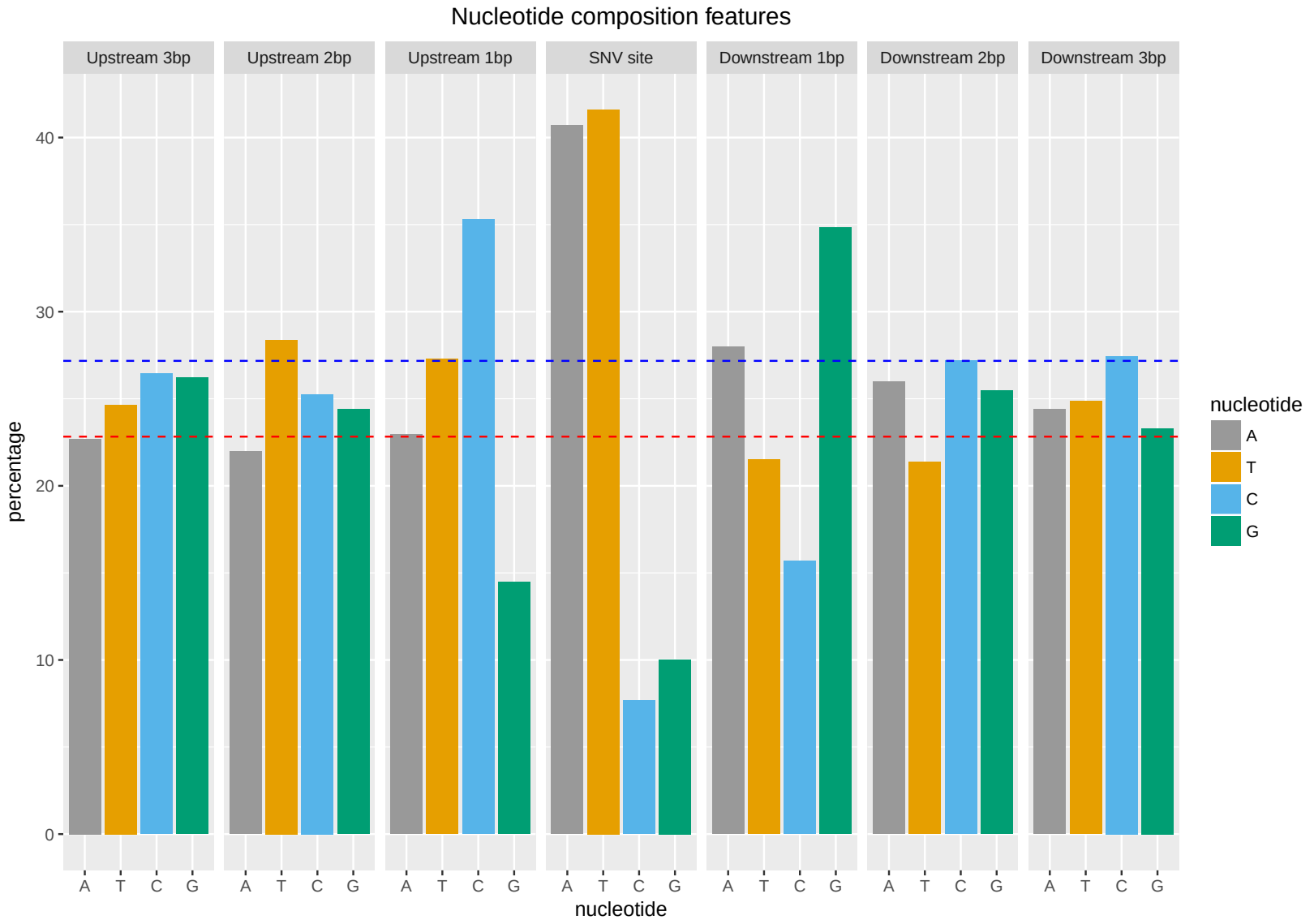

Nucleotide composition features
Upstream 3bp
Upstream 2bp
Upstream 1bp
SNV site
Downstream 1bp
Downstream 2bp
Downstream 3bp
percentage
0
10
20
30
40
A
T
C
G
nucleotide
nucleotide
A
T
C
G

Figure S7 part 1

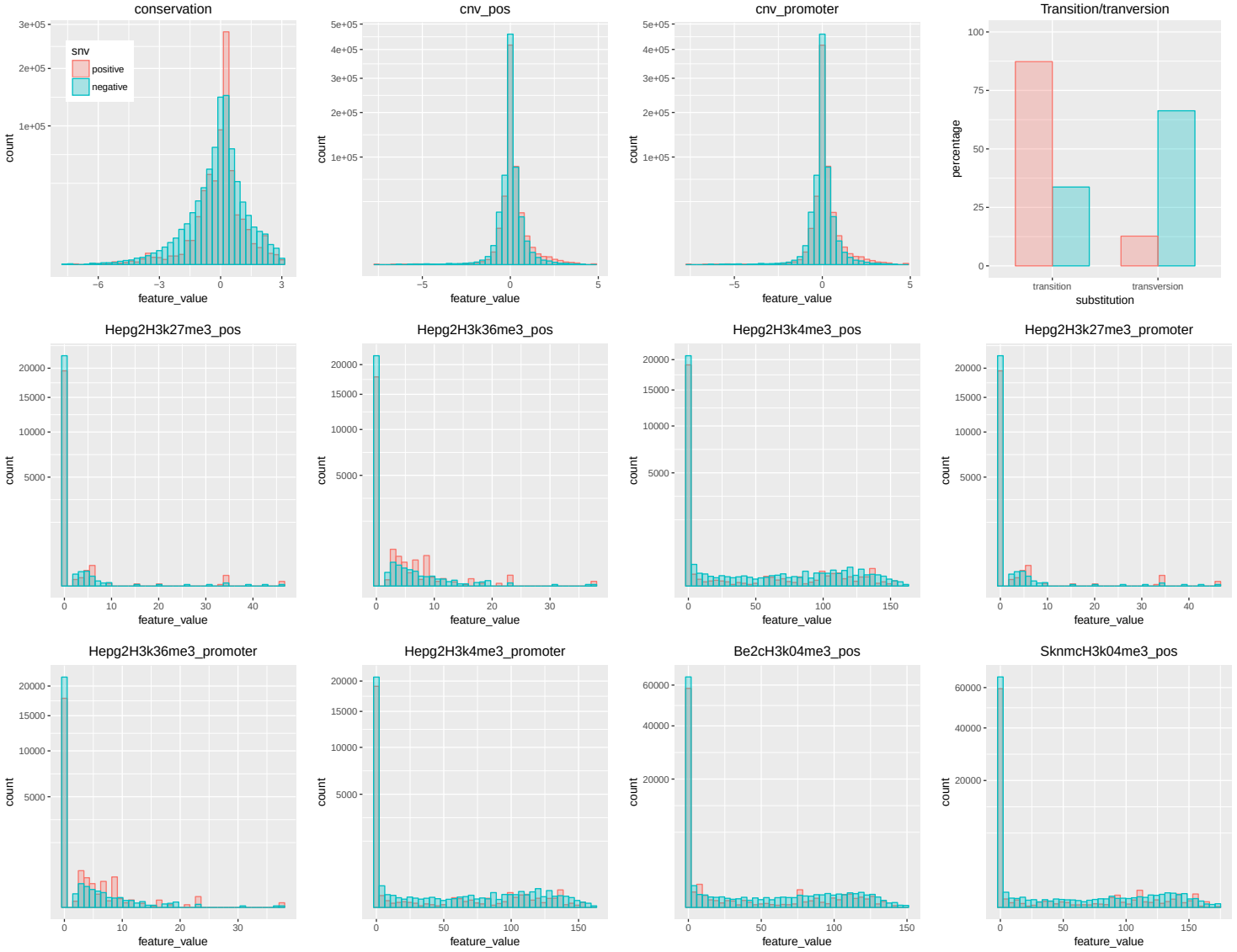


Figure S7 part 2

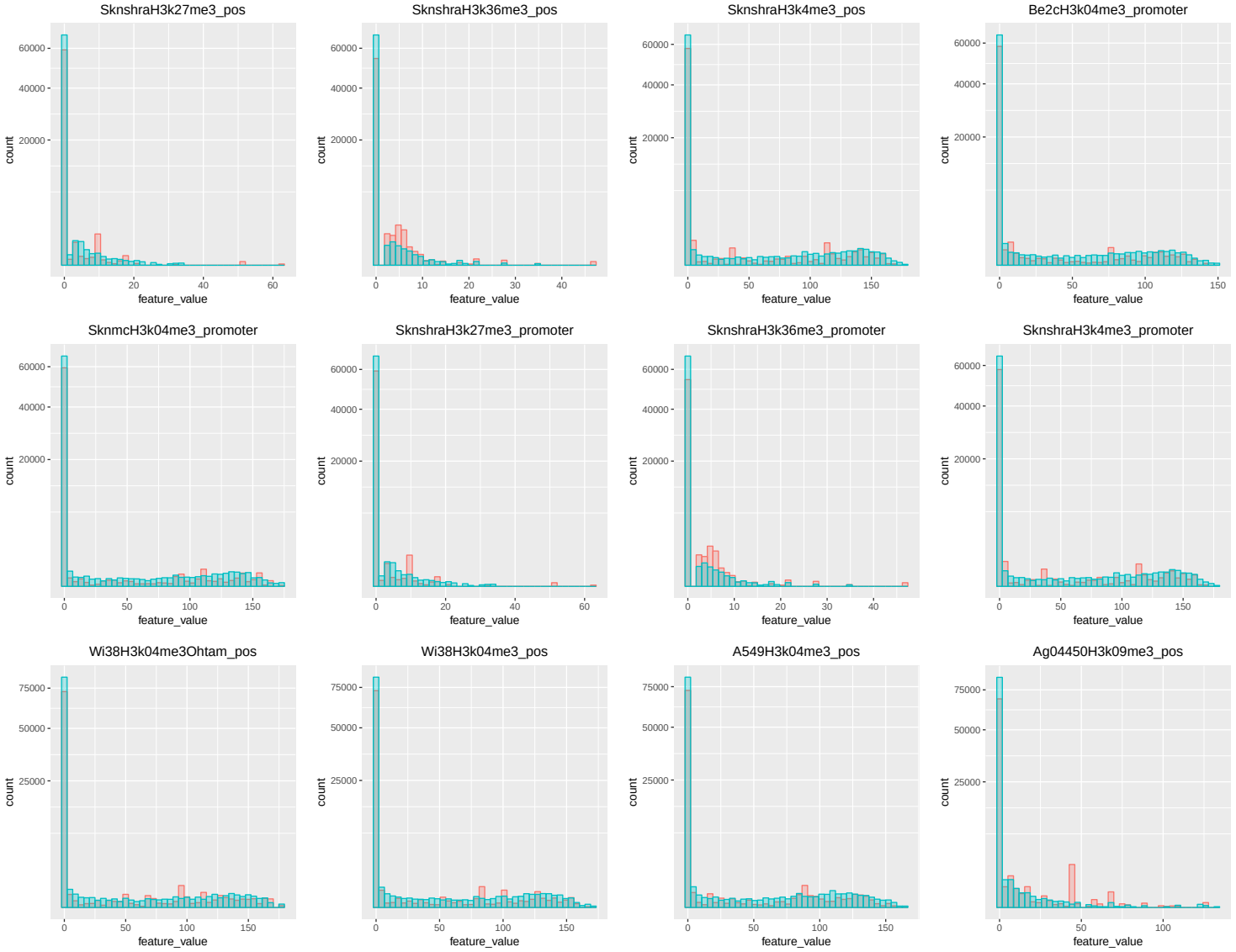


Figure S7 part 3

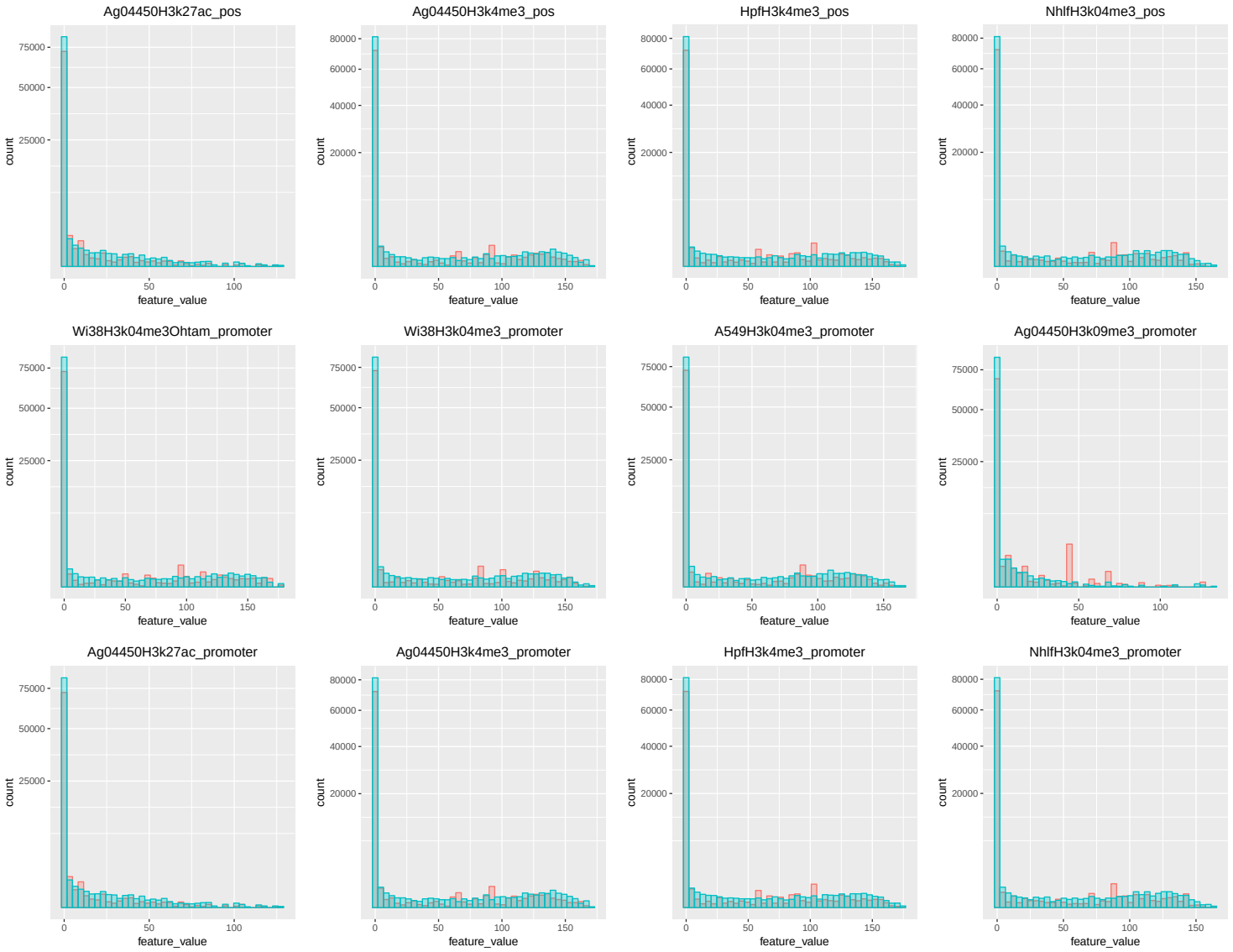


Figure S7 part 4

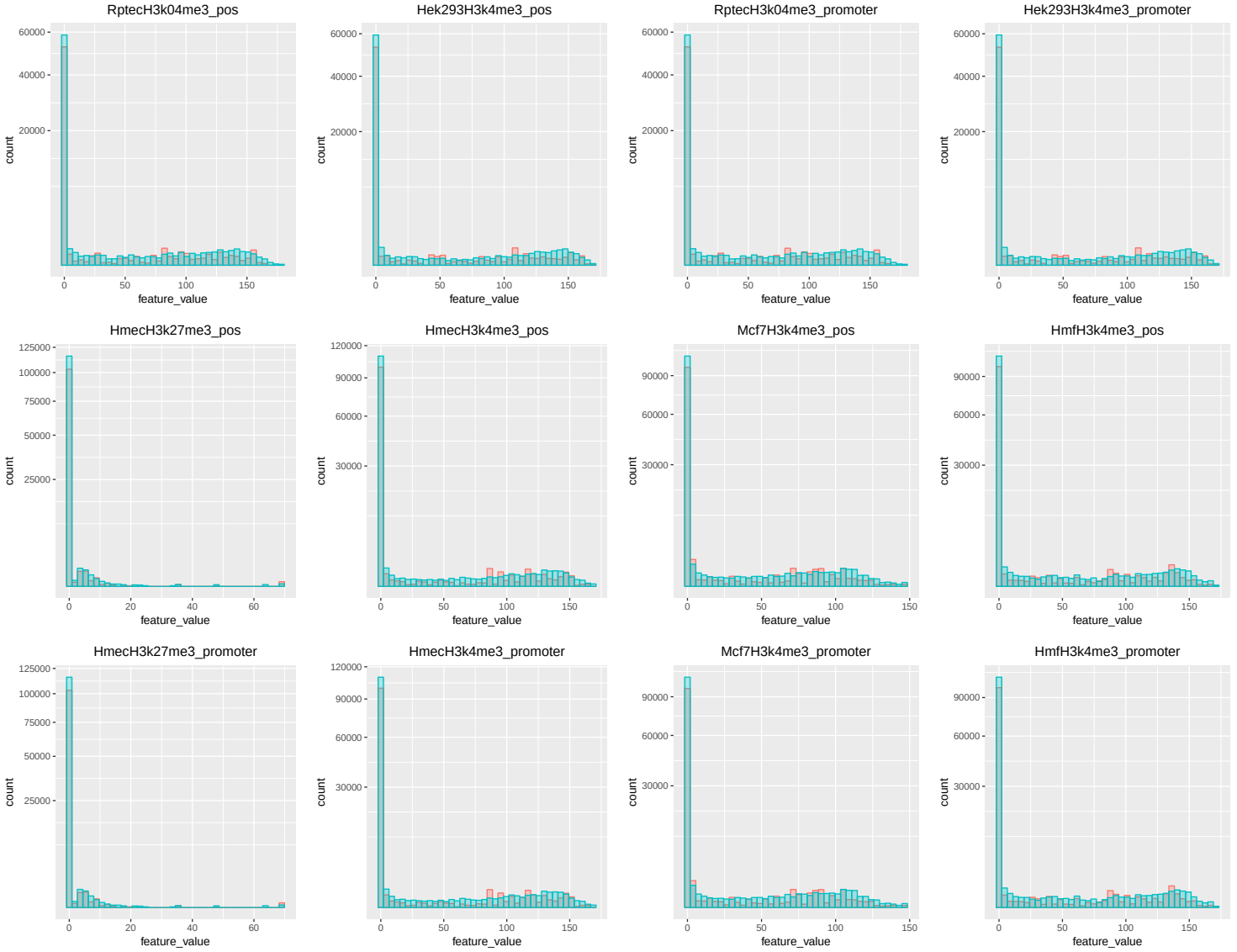


Figure S8

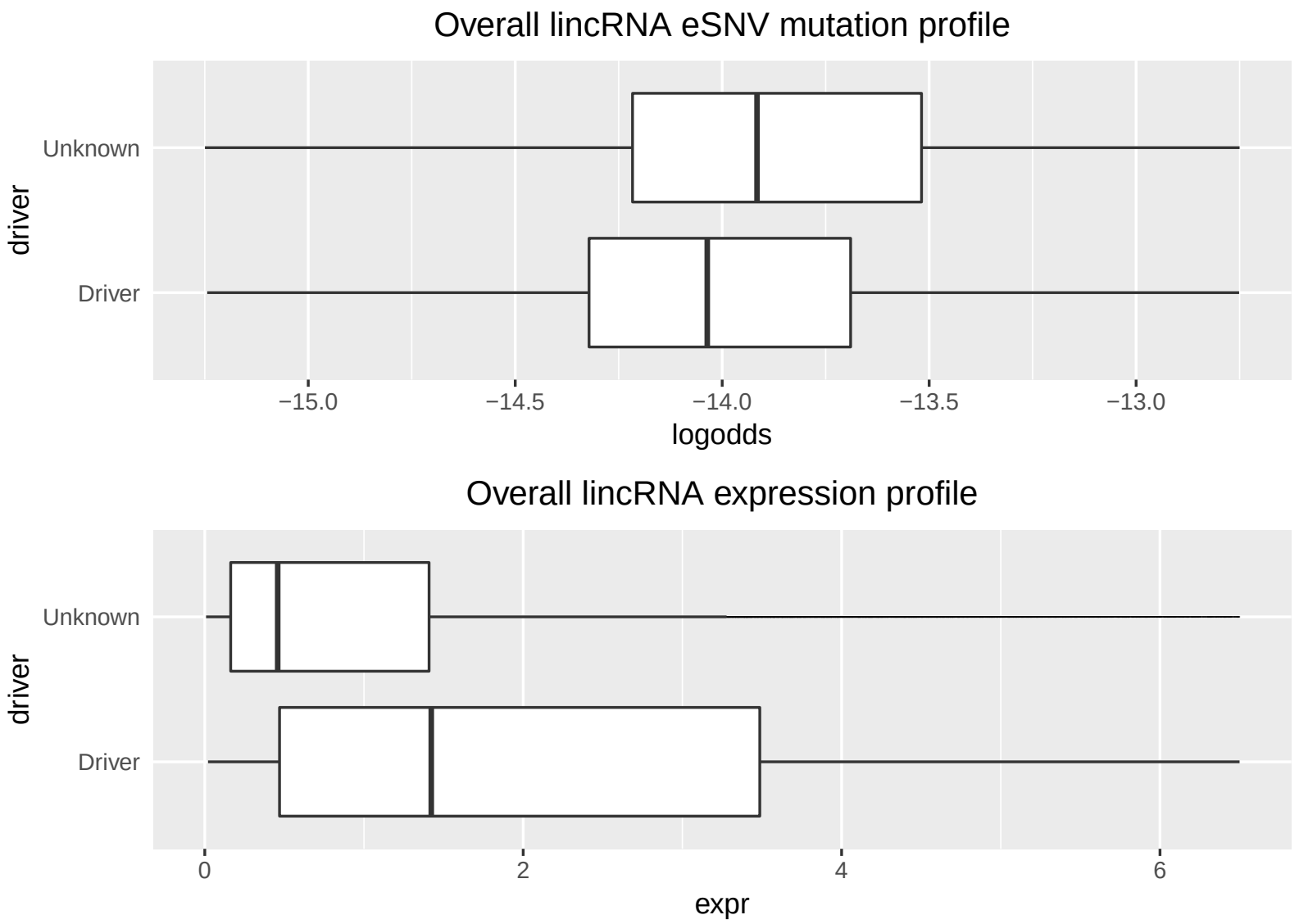

Overall lincRNA eSNV mutation profile
driver
Unknown
Driver
−15.0
−14.5
−14.0
−13.5
−13.0
logodds
Overall lincRNA expression profile
driver
Unknown
Driver
0
2
4
6
expr

Figure S9

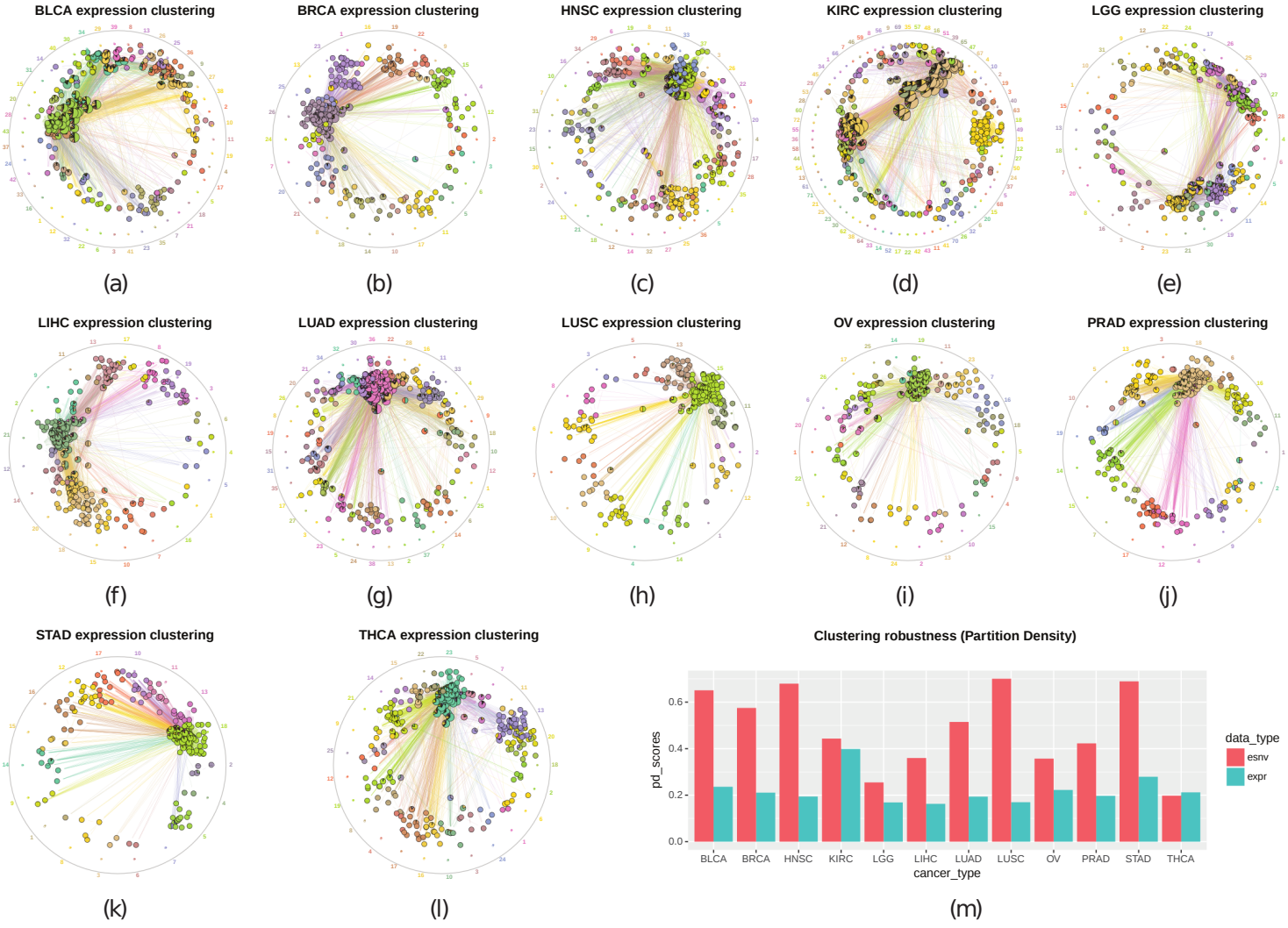
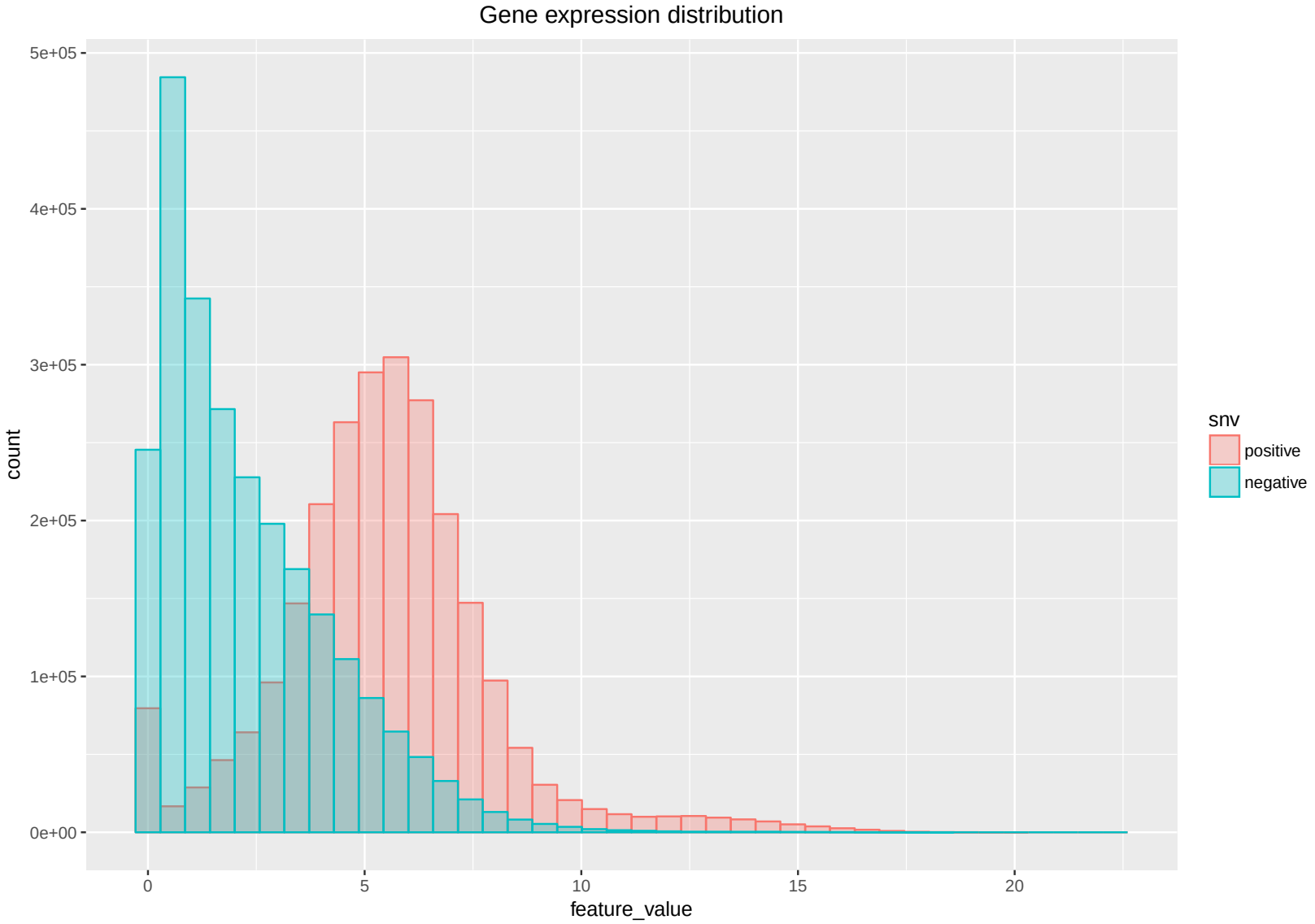

BLCA expression clustering
BRCA expression clustering
HNSC expression clustering
KIRC expression clustering
LGG expression clustering
(a)
(b)
(c)
(d)
(e)
LIHC expression clustering
LUAD expression clustering
LUSC expression clustering
OV expression clustering
PRAD expression clustering
(f)
(g)
(h)
(i)
(j)
STAD expression clustering
THCA expression clustering
Clustering robustness (Partition Density)
pd_scores
0.0
0.2
0.4
0.6
BLCA
BRCA
HNSC
KIRC
LGG
LIHC
LUAD
LUSC
OV
PRAD
STAD
THCA
cancer_type
data_type
esnv
expr
(k)
(l)
(m)

Figure S10


Gene expression distribution
count
5e+05
4e+05
3e+05
2e+05
1e+05
0e+00
0
5
10
15
20
feature_value
snv
positive
negative

6.6.2 Supplemental information

Supplementary Methods

Software versions used

- Java runtime environment: version 1.8.0_66
- STAR: version 2.5.1b
- GenomeAnalysisTK: version 3.5-0-g36282e4
- Picard Tools: version 2.1.1

Reference files used:

- Genome hg19 (Illumina iGenomes) downloaded from Illumina:
- dbSNP (version 138), Mills/1000 Genomes gold standard indels, 1000 Genomes downloaded from the Broad Institute: phase1 indels

eSNV processing steps:

1. Fastq files were aligned using the STAR aligner using 20 threads, using "basic" twopassMode
2. Read groups were added to each sample using picard tools AddOrReplaceReadGroups function
3. Duplicates were handled using picard tools MarkDupliactes function with parameters:
    - -CREATE_INDEX=true
    - VALIDATION_STRINGENCY=SILENT
4. Cigar strings were split using GenomeAnalysisTK SplitNCigarReads function with parameters:
    - -rf ReassignOneMappingQuality

- -RMQF 255
- -RMQT 60
- -U ALLOW_N_CIGAR_READS

5. Indel realignment in GenomeAnalysisTK was performed using the Mills/1000 Genomes gold standard and phase 1 indels as additional references. The RealignerTargetCreator function was used followed by the IndelRealigner function.

6. Base recalibration in GenomeAnalysisTK was performed using the dbSNP, Mills/1000 Genomes gold standard and phase 1 indels as additional references. The BaseRecalibrator function was used followed by the PrintReads function.

7. Raw variant calling in GenomeAnalysisTK was performed using dbSNP variants as additional references. The HaplotypeCaller function was used with the following parameters:

   - -dontUseSoftClippedBases
   - -stand_call_conf 20.0

8. -stand_emit_conf 20.0

9. -o $analysis_id.raw.vcf

10. 8) Additional filtration was performed on the called variants using the VariantFiltration function. Additional filtration parameters used were:

11. -window 35

12. -cluster 3

    - -filterName FS
    - -filter "FS >30.0"
    - -filterName QD
    - -filter "QD <2.0"

6.7 Chapter summary

We use expressed single nucleotide variants (eSNV) in lincRNAs from RNA-Seq data to determine important relations between somatic mutations and other molecular and clinical data. Over 6000 RNA-Seq samples from 12 cancer types were processed for both eSNV data and lincRNA and gene expression data. We use a variety of machine learning methods to determine the molecular features that are highly correlated with somatic mutations. Subsequently, we perform factor correlation using mutual information to find which molecular features cluster together.

By asking what are the fundamental biological and molecular properties that are attributed to somatic lincRNA mutations in cancer, this may eventually lead to a more fundamental understanding of the tumorigenic process of genetic mutations leading to malignancy.

Chapter 7

Discussion

After the human genome project was completed in 2003, many scientists believed that we would be able to soon fully understand our human biological nature and inner workings of our cells. In fact, the original goal of the Human Genome Project was "the complete mapping and understanding of all the genes of human beings." [1] While the first goal, of mapping the genome, has been thoroughly fulfilled, the second goal, of understanding all the genes of human beings, will not be reasonably met in the near future.

While we know a relatively well the 2-3% of the genome that produces proteins and enzymes, we know very little about majority of the 80% of the genome that is transcribed. This genomic "dark matter" is profoundly unknown, and we are only just beginning to understand the functions and processes of these transcribed non-coding regions. The study of non-coding RNAs is still in its infancy.

Early genetic experiments on bacteria and single-celled systems, with limited ability to characterize the full extent of DNA and RNA material, focused on abundant protein structures and defined our understanding of the biological centra dogma – DNA is transcribed to RNA is transcribed to proteins [2]. In simple organisms such as bacteria, lincRNAs and other non-coding RNA elements generally do not exist.

As early as 2004, researchers found evidence that our basic understanding of how cells work was fundamentally flawed. In an opinion article in Nature Genetics [3], John S. Mattick noted that in bacteria, the number of regulatory proteins compared with the number of enzymatic proteins had a quadratic relationship. But in eukaryotes this trend did not continue. Thus, extrapolating from our prokaryotic counterparts, there clearly seemed to be a huge amount of missing regulatory proteins that would be required for our cells to operate in a cohesive manner.

Mattick suggested that the large deficiency in the number of expected regulatory proteins was made up by non-coding RNAs – something that was missed completely in the original annotation of the human genome. With the discovery of thousands of robustly transcribed non-coding RNA through high-throughput sequencing, Mattick seems to have been proven correct. Indeed, even among non-coding RNA species, the intergenic lincRNAs subclass is the largest group of RNA transcripts found in the entire human genome [4].

7.1 Completion of specific aims

At the beginning on my doctoral project, I research the current state of the art in lincRNAs and next generation sequencing. In the lincRNA review paper (Chapter 2), I comprehensively explore the lincRNA literature, and focus on learning how lincRNAs are connected to cancer, existing computational methods related to lincRNA research, and the challenges in this field. To become proficient in analyzing RNA-Seq data, I perform a meta-analysis on differential expression methods (Chapter 3). Using several public datasets, I evaluate the performance of various differential expression tools, and study the relationship of sample size and sequencing depth to the statistical detection power of different experimental designs.

In Aim 1 of this project, I process thousands of RNA-Seq cancer and tumor-adjacent normal samples and quantify the expression of lincRNAs in these samples (Chapter 4). I explor the expression landscape of lincRNAs across 12 cancer types, and find six common lincRNAs that were differentially expressed in all cancer types. Using machine learning methods, the expression of these lincRNAs can differentiate tumor and normal tissue samples with near-perfect accuracy. Furthermore, these six lincRNAs are significantly correlated with patient survival in several cancer datasets.

In Chapter 5, I explor how neural networks could be used to analyze high throughput datasets linked to patient survival. I show that neural network survival analysis, termed Cox-nnet, also has a unique advantage over other machine learning methods, such as revealing the biological relevance of genes and pathways correlated with survival.

In aim 2 of my dissertation, I explore the mutational landscape of lincRNAs (Chapter 6). Specifically, I analyze the expressed single nucleotide variations in cancer. Computationally, this is my most ambitious project yet, and one that I would not have been able to accomplish at the start of my research. This paper required large scale data processing on the UH manoa high performance computing. From 6000+ RNA-Seq samples, I extract features on over 300 million single nucleotide variations and leverage state of the art machine learning methods. For

the result of this paper, I determine which features cause nucleotide positions to be more or less likely to be mutated in cancer. I was also able to show that lincRNAs which are known to be cancer drivers have different mutation and expression profiles.

7.2 Future work and directions

Previous GWAS studies have also linked disease-associated genetic variation and somatic mutations to regions inside of, or in the vicinity of lincRNAs. LincRNAs are expressed differentially in not only different tissues, but in different cancer types and subtypes. Thus, exploration of the lincRNA landscape across cancers should focus on not only the common lincRNAs, but also incorporate tissue and cancer specificity in analyzing the expression and mutation profiles.

To follow up on Aim 1, it will be necessary to fully sequence the six pan-cancer biomarkers. Currently, the isoforms and exon structure of these lincRNAs are only computationally predicted through high-throughput sequencing. Rapid amplification of cDNA ends (RACE) PCR will helpful to explicitly determine the isoform structures and sequences. Further work should be performed to validate the potential clinical applications of these biomarkers. Additional tissues and also blood serum on retrospective and prospective cohorts will be great resources to further explore the expression and presence of these lincRNAs as biomarkers. In addition, it might be interesting to determine the molecular function and biological pathways that these lincRNAs are involved in.

To follow up on Aim 2, somatic mutations from whole genome sequencing should be processed. While whole genome sequencing processed results were not available at the start of this project, WGS should provide complementary data to the eSNV data from RNA-Seq. Allele specific expression from the RNA-Seq data could be an important factor in the analysis. Somatic mutations in promoter and enhancer regions determined by WGS could complement the analysis. Further computational work should also be performed in order to predict if a mutation on a given nucleotide postion would have a deleterious effect or any disease association in cancer.

Bibliography

1. Aban, I. B., Cutter, G. R. & Mavinga, N. Inferences and Power Analysis Concerning Two Negative Binomial Distributions with An Application to MRI Lesion Counts Data. *Computational statistics & data analysis* **53,** 820–833. ISSN: 0167-9473 (2008).
2. Ahn, Y.-Y., Bagrow, J. P. & Lehmann, S. Link communities reveal multiscale complexity in networks. *Nature* **466,** 761–764 (2010).
3. Akhurst, R. J. & Derynck, R. TGF-Beta signaling in cancer-a double-edged sword. *Trends in cell biology* **11,** S44–S51. ISSN: 0962-8924 (2001).
4. Al-Rfou, R., Alain, G., Almahairi, A., Angermueller, C., Bahdanau, D., Ballas, N., Bastien, F., Bayer, J., Belikov, A. & Belopolsky, A. Theano: A Python framework for fast computation of mathematical expressions. *arXiv preprint arXiv:1605.02688* (2016).
5. Al-Tassan, N. A., Whiffin, N., Hosking, F. J., Palles, C., Farrington, S. M., Dobbins, S. E., Harris, R., Gorman, M., Tenesa, A. & Meyer, B. F. A new GWAS and meta-analysis with 1000Genomes imputation identifies novel risk variants for colorectal cancer. *Scientific reports* **5** (2015).
6. Anders, S. Analysing RNA-Seq data with the DESeq package. *Mol Biol,* 1–17 (2010).
7. Anders, S. HTSeq: Analysing high-throughput sequencing data with Python. *Bioinformatics* (2010).
8. Anders, S. & Huber, W. Differential expression analysis for sequence count data. *Genome Biol* **11** (2010).
9. Bansal, M., Belcastro, V., Ambesi-Impiombato, A. & Di Bernardo, D. How to infer gene networks from expression profiles. *Molecular systems biology* **3.** ISSN: 1744-4292 (2007).
10. Bengio, Y., Boulanger-Lewandowski, N. & Pascanu, R. *Advances in optimizing recurrent networks* in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on* (IEEE, 2013), 8624–8628. ISBN: 1520-6149.

11. Berrar, D., Bradbury, I. & Dubitzky, W. Avoiding model selection bias in small-sample genomic datasets. *Bioinformatics* **22,** 1245–1250. ISSN: 1367-4803 (2006).

12. Bi, R., Liu, P. & Triche, T. Package 'ssizeRNA'. *Bioconductpor* (2016).

13. Binder, H. CoxBoost: Cox models by likelihood based boosting for a single survival endpoint or competing risks. *R package version* **1** (2013).

14. Boerner, S. & McGinnis, K. M. Computational identification and functional predictions of long noncoding RNA in Zea mays. *PloS one* **7,** e43047. ISSN: 1932-6203 (2012).

15. Bottomly, D., Walter, N. A., Hunter, J. E., Darakjian, P., Kawane, S., Buck, K. J., Searles, R. P., Mooney, M., McWeeney, S. K. & Hitzemann, R. Evaluating gene expression in C57BL/6J and DBA/2J mouse striatum using RNA-Seq and microarrays. *PLoS One* **6,** e17820. ISSN: 1932-6203 (2011).

16. Breheny, P. & Huang, J. Coordinate descent algorithms for nonconvex penalized regression, with applications to biological feature selection. *The annals of applied statistics* **5,** 232 (2011).

17. Broad. Broad Institute TCGA Genome Data Analysis Center (2014): Analysis Overview for 15 July 2014. *Broad Institute of MIT and Harvard.* doi:10.7908/C1DN43V9 (2014).

18. Brockdorff, N., Ashworth, A., Kay, G. F., Cooper, P., Smith, S., McCabe, V. M., Norris, D. P., Penny, G. D., Patel, D. & Rastan, S. Conservation of position and exclusive expression of mouse Xist from the inactive X chromosome. *Nature* **351,** 329–331. ISSN: 0028-0836 (1991).

19. Bryois, J., Buil, A., Evans, D. M., Kemp, J. P., Montgomery, S. B., Conrad, D. F., Ho, K. M., Ring, S., Hurles, M., Deloukas, P. & et al. Cis and trans effects of human genomic variants on gene expression. *PLoS Genet* **10,** e1004461 (2014).

20. Bullard, J. H., Purdom, E., Hansen, K. D. & Dudoit, S. Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. *BMC bioinformatics* **11.** ISSN: 1471-2105. doi:10.1186/1471-2105-11-94 (2010).

21. Busby, M. A., Stewart, C., Miller, C. A., Grzeda, K. R. & Marth, G. T. Scotty: a web tool for designing RNA-Seq experiments to measure differential gene expression. *Bioinformatics* **29,** 656–657 (2013).

22. Cabili, M. N., Trapnell, C., Goff, L., Koziol, M., Tazon-Vega, B., Regev, A. & Rinn, J. L. Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. *Genes & Development* **25,** 1915–1927. ISSN: 0890-9369, 1549-5477 (2011).

23. Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K. & Madden, T. L. BLAST+: architecture and applications. *BMC bioinformatics* **10,** 421. ISSN: 1471-2105 (2009).

24. Cancer Genome Atlas Research, N., Weinstein, J. N., Collisson, E. A., Mills, G. B., Shaw, K. R., Ozenberger, B. A., Ellrott, K., Shmulevich, I., Sander, C. & Stuart, J. M. The Cancer Genome Atlas Pan-Cancer analysis project. *Nat Genet* **45,** 1113–20. ISSN: 1546-1718 (Electronic) 1061-4036 (Linking) (2013).

25. Cancer Genome Atlas, N. Comprehensive molecular portraits of human breast tumours. *Nature* **490,** 61–70. ISSN: 1476-4687 (Electronic) 0028-0836 (Linking) (2012).

26. Chen, T. & Guestrin, C. Xgboost: A scalable tree boosting system. *Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining,* 785–794 (2016).

27. Chen, Z., Liu, J., Ng, H. K. T., Nadarajah, S., Kaufman, H. L., Yang, J. Y. & Deng, Y. Statistical methods on detecting differentially expressed genes for RNA-seq data. *BMC Systems Biology* **5** (2011).

28. Chi, C.-L., Street, W. N. & Wolberg, W. H. Application of artificial neural network-based survival analysis on two breast cancer datasets. *AMIA Symposium* **2007,** 130 (2007).

29. Ching, T. & Garmire, L. X. Pan-cancer analysis of expressed single nucleotide variants in long intergenic non-coding RNA. *(under review)* (2017).

30. Ching, T., Huang, S. & Garmire, L. X. Power analysis and sample size estimation for RNA-Seq differential expression. *Rna* **20,** 1684–1696. ISSN: 1469-9001 (Electronic) 1355-8382 (Linking) (2014).

31. Ching, T., Masaki, J., Weirather, J. & Garmire, L. X. Non-coding yet non-trivial: a review on the computational genomics of lincRNAs. *BioData mining* **8,** 44 (2015).

32. Ching, T., Peplowska, K., Huang, S., Zhu, X., Shen, Y., Molnar, J., Yu, H., Tiirikainen, M., Fogelgren, B., Fan, R. & Garmire, L. X. Pan-cancer analyses reveal long intergenic non-coding RNAs relevant to tumor diagnosis, subtyping and prognosis. *EBioMedicine* **7,** 62–72 (2016).

33. Ching, T., Zhu, X. & Garmire, L. X. Cox-nnet: a neural network extension to Cox regression. *(under review)* (2017).

34. Chiyomaru, T., Fukuhara, S., Saini, S., Majid, S., Deng, G., Shahryari, V., Chang, I., Tanaka, Y., Enokida, H. & Nakagawa, M. Long non-coding RNA HOTAIR is targeted and regulated by miR-141 in human cancer cells. *Journal of Biological Chemistry* **289,** 12550–12565. ISSN: 0021-9258 (2014).

35. Choueiri, T. K., Vaishampayan, U., Rosenberg, J. E., Logan, T. F., Harzstark, A. L., Bukowski, R. M., Rini, B. I., Srinivas, S., Stein, M. N. & Adams, L. M. Phase II and biomarker study of the dual MET/VEGFR2 inhibitor foretinib in patients with papillary renal cell carcinoma. *Journal of Clinical Oncology* **31,** 181–186. ISSN: 0732-183X (2013).

36. Cibulskis, K., Lawrence, M. S., Carter, S. L., Sivachenko, A., Jaffe, D., Sougnez, C., Gabriel, S., Meyerson, M., Lander, E. S. & Getz, G. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nature biotechnology* **31,** 213–219 (2013).

37. Consortium, E. P. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489,** 57–74. ISSN: 0028-0836 (2012).

38. Consortium, E. P. *et al.* The ENCODE (ENCyclopedia of DNA elements) project. *Science* **306,** 636–640 (2004).

39. Cork, S. M. & Van Meir, E. G. Emerging roles for the BAI1 protein family in the regulation of phagocytosis, synaptogenesis, neurovasculature, and tumor development. *Journal of molecular medicine* **89,** 743–752. ISSN: 0946-2716 (2011).

40. Cornish, A. & Guda, C. A comparison of variant calling pipelines using genome in a bottle as a reference. *BioMed research international* **2015** (2015).

41. Costello, M., Pugh, T. J., Fennell, T. J., Stewart, C., Lichtenstein, L., Meldrim, J. C., Fostel, J. L., Friedrich, D. C., Perrin, D., Dionne, D. & et al. Discovery and characterization of artifactual mutations in deep coverage targeted capture sequencing data due to oxidative DNA damage during sample preparation. *Nucleic acids research,* gks1443 (2013).

42. Cover, T. M. & Thomas, J. A. *Elements of information theory* (John Wiley & Sons, 2012).

43. Crick, F. H. *On protein synthesis* in *Symp Soc Exp Biol* **12** (1958), 8.

44. Dal Pozzolo, A., Caelen, O., Johnson, R. A. & Bontempi, G. Calibrating Probability with Undersampling for Unbalanced Classification. *Computational Intelligence, 2015 IEEE Symposium Series,* 159–166 (2015).

45. Demuth, H. B., Beale, M. H., De Jess, O. & Hagan, M. T. *Neural network design* ISBN: 0971732116 (Martin Hagan, 2014).

46. Di Croce, L. & Helin, K. Transcriptional regulation by Polycomb group proteins. *Nat Struct Mol Biol* **20,** 1147–55. ISSN: 1545-9985 (Electronic) 1545-9985 (Linking) (2013).

47. Dimitrova, N., Zamudio, J. R., Jong, R. M., Soukup, D., Resnick, R., Sarma, K., Ward, A. J., Raj, A., Lee, J. T., Sharp, P. A. & Jacks, T. LincRNA-p21 activates p21 in cis to promote Polycomb target gene expression and to enforce the G1/S checkpoint. *Mol Cell* **54,** 777–90. ISSN: 1097-4164 (Electronic) 1097-2765 (Linking) (2014).

48. Ding, L., Getz, G., Wheeler, D. A., Mardis, E. R., McLellan, M. D., Cibulskis, K., Sougnez, C., Greulich, H., Muzny, D. M., Morgan, M. B. & et al. Somatic mutations affect key pathways in lung adenocarcinoma. *Nature* **455,** 1069–1075 (2008).

49. Djebali, S. *et al.* Landscape of transcription in human cells. *Nature* **489,** 101–8. ISSN: 1476-4687 (Electronic) 0028-0836 (Linking) (2012).

50. Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M. & Gingeras, T. R. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29,** 15–21 (2013).

51. Donaldson, J. & Donaldson, M. J. Package 'tsne'. *CRAN Repository* (2010).

52. Du, Z., Fei, T., Verhaak, R. G., Su, Z., Zhang, Y., Brown, M., Chen, Y. & Liu, X. S. Integrative genomic analyses reveal clinically relevant long noncoding RNAs in human cancer. *Nat Struct Mol Biol* **20,** 908–13. ISSN: 1545-9985 (Electronic) 1545-9985 (Linking) (2013).

53. Dunham, I. *et al.* An integrated encyclopedia of DNA elements in the human genome. *Nature* **489,** 57–74. ISSN: 1476-4687 (Electronic) 0028-0836 (Linking) (2012).

54. Durinck, S., Spellman, P. T., Birney, E. & Huber, W. Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. *Nature protocols* **4,** 1184–1191 (2009).

55. Duss, O., Michel, E., Yulikov, M., Schubert, M., Jeschke, G. & Allain, F. H. T. Structural basis of the non-coding RNA RsmZ acting as a protein sponge. *Nature* **509,** 588–+. ISSN: 0028-0836 (2014).

56. ENCODE. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489,** 57–74. ISSN: 0028-0836 (2012).

57. Ezkurdia, I., Juan, D., Rodriguez, J. M., Frankish, A., Diekhans, M., Harrow, J., Vazquez, J., Valencia, A. & Tress, M. L. Multiple evidence strands suggest that there may be as few as 19 000 human protein-coding genes. *Hum Mol Genet.* ISSN: 1460-2083 (Electronic) 0964-6906 (Linking). doi:10.1093/hmg/ddu309 (2014).

58. Fan, X.-N. & Zhang, S.-W. lncRNA-MFDL: identification of human long non-coding RNAs by fusing multiple features and using deep learning. *Molecular BioSystems* (2015).

59. Faraggi, D. & Simon, R. A neural network model for survival data. *Statistics in medicine* **14,** 73–82. ISSN: 1097-0258 (1995).

60. Flicek, P. *et al.* Ensembl 2012. *Nucleic Acids Res* **40,** D84–90. ISSN: 1362-4962 (Electronic) 0305-1048 (Linking) (2012).

61. Frazee, A., Langmead, B. & Leek, J. ReCount: A multi-experiment resource of analysis-ready RNA-seq gene count datasets. *BMC bioinformatics* **12,** 449. ISSN: 1471-2105 (2011).

62. Friday, B. B. & Adjei, A. A. Advances in targeting the Ras/Raf/MEK/Erk mitogen-activated protein kinase cascade with MEK inhibitors for cancer therapy. *Clinical Cancer Research* **14,** 342–346. ISSN: 1078-0432 (2008).

63. Fukushima, Y., Oshika, Y., Tsuchida, T., Tokunaga, T., Hatanaka, H., Kijima, H., Yamazaki, H., Ueyama, Y., Tamaoki, N. & Nakamura, M. Brain-specific angiogenesis inhibitor 1 expression is inversely correlated with vascularity and distant metastasis of colorectal cancer. *International journal of oncology* **13,** 967–970. ISSN: 1019-6439 (1998).

64. Galperin, M. Y., Rigden, D. J. & Fernandez-Suarez, X. M. The 2015 Nucleic Acids Research Database Issue and Molecular Biology Database Collection. *Nucleic acids research* **43,** D1–D5. ISSN: 0305-1048 (2015).

65. Garmire, L. X., Garmire, D. G., Huang, W., Yao, J., Glass, C. K. & Subramaniam, S. A global clustering algorithm to identify long intergenic non-coding RNA-with applications in mouse macrophages. *PLoS One* **6,** e24051. ISSN: 1932-6203 (2011).

66. Ge, X., Chen, Y., Liao, X., Liu, D., Li, F., Ruan, H. & Jia, W. Overexpression of long noncoding RNA PCAT-1 is a novel biomarker of poor prognosis in patients with colorectal cancer. *Medical oncology* **30,** 1–6. ISSN: 1357-0560 (2013).

67. Gerds, T. A., Kattan, M. W., Schumacher, M. & Yu, C. Estimating a time-dependent concordance index for survival prediction models with covariate dependent censoring. *Statistics in Medicine* **32,** 2173–2184. ISSN: 1097-0258 (2013).

68. Gevrey, M., Dimopoulos, I. & Lek, S. Review and comparison of methods to study the contribution of variables in artificial neural network models. *Ecological modelling* **160,** 249–264. ISSN: 0304-3800 (2003).

69. Girgin, C., Tarhan, H., seyin uuml, u., Hekimgil, M., Sezer, A. & Gurel, G. P53 mutations and other prognostic factors of renal cell carcinoma. *Urologia internationalis* **66,** 78–83. ISSN: 1423-0399 (2001).

70. Glazko, G. V., Zybailov, B. L. & Rogozin, I. B. Computational prediction of polycomb-associated long non-coding RNAs. *PloS one* **7,** e44878. ISSN: 1932-6203 (2012).

71. Goff, L. A. & Rinn, J. L. Poly-combing the genome for RNA. *Nature structural & molecular biology* **20,** 1344–1346. ISSN: 1545-9993 (2013).

72. Gong, C. & Maquat, L. E. *Affinity Purification of Long Noncoding RNA-Protein Complexes from Formaldehyde Cross-Linked Mammalian Cells* 81–86. ISBN: 1493913689 (Springer, 2015).

73. Gonzalez-Perez, A., Perez-Llamas, C., Deu-Pons, J., Tamborero, D., Schroeder, M. P., Jene-Sanz, A., Santos, A. & Lopez-Bigas, N. IntOGen-mutations identifies cancer drivers across tumor types. *Nature methods* **10,** 1081–1082 (2013).

74. Griffiths-Jones, S., Moxon, S., Marshall, M., Khanna, A., Eddy, S. R. & Bateman, A. Rfam: annotating non-coding RNAs in complete genomes. *Nucleic acids research* **33,** D121–D124. ISSN: 0305-1048 (2005).

75. Guo, X., Gao, L., Liao, Q., Xiao, H., Ma, X., Yang, X., Luo, H., Zhao, G., Bu, D. & Jiao, F. Long non-coding RNAs function annotation: a global prediction method based on bi-colored networks. *Nucleic acids research* **41,** e35–e35. ISSN: 0305-1048 (2013).

76. Gupta, R. A., Shah, N., Wang, K. C., Kim, J., Horlings, H. M., Wong, D. J., Tsai, M.-C., Hung, T., Argani, P. & Rinn, J. L. Long non-coding RNA HOTAIR reprograms chromatin state to promote cancer metastasis. *Nature* **464,** 1071–1076. ISSN: 0028-0836 (2010).

77. Gutschner, T. & Diederichs, S. The hallmarks of cancer: a long non-coding RNA point of view. *RNA Biol* **9,** 703–19. ISSN: 1555-8584 (Electronic) 1547-6286 (Linking) (2012).

78. Guttman, M. *et al.* Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. *Nature* **458,** 223–7. ISSN: 1476-4687 (Electronic) 0028-0836 (Linking) (2009).

79. Guttman, M. *et al.* Ab initio reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs. *Nature Biotechnology* **28,** 503–U166. ISSN: 1087-0156 (2010).

80. Guttman, M., Russell, P., Ingolia, N. T., Weissman, J. S. & Lander, E. S. Ribosome profiling provides evidence that large noncoding RNAs do not encode proteins. *Cell* **154,** 240–251. ISSN: 0092-8674 (2013).

81. Habel, L. A., Shak, S., Jacobs, M. K., Capra, A., Alexander, C., Pho, M., Baker, J., Walker, M., Watson, D. & Hackett, J. A population-based study of tumor gene expression and risk of breast cancer death among lymph node-negative patients. *Breast Cancer Res* **8,** R25 (2006).

82. Han, L., Yuan, Y., Zheng, S., Yang, Y., Li, J., Edgerton, M. E., Diao, L., Xu, Y., Verhaak, R. G. & Liang, H. The Pan-Cancer analysis of pseudogene expression reveals biologically and clinically relevant tumour subtypes. *Nature communications* **5** (2014).

83. Hangauer, M. J., Vaughn, I. W. & McManus, M. T. Pervasive transcription of the human genome produces thousands of previously unidentified long intergenic noncoding RNAs. *PLoS genetics* **9,** e1003569. ISSN: 1553-7404 (2013).

84. Harrell, F. E., Lee, K. L. & Mark, D. B. Tutorial in biostatistics multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Statistics in medicine* **15,** 361–387. ISSN: 0277-6715 (1996).

85. Harrow, J. *et al.* GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Res* **22,** 1760–74. ISSN: 1549-5469 (Electronic) 1088-9051 (Linking) (2012).

86. Hart, S. N., Therneau, T. M., Zhang, Y., Poland, G. A. & Kocher, J.-P. Calculating Sample Size Estimates for RNA Sequencing Data. *Journal of Computational Biology* **20,** 970–978. ISSN: 1066-5277 (2013).

87. He, Q., He, Q., Liu, X., Wei, Y., Shen, S., Hu, X., Li, Q., Peng, X., Wang, L. & Yu, L. Genome-wide prediction of cancer driver genes based on SNP and cancer SNV data. *American journal of cancer research* **4,** 394 (2014).

88. He, Y., Vogelstein, B., Velculescu, V. E., Papadopoulos, N. & Kinzler, K. W. The antisense transcriptomes of human cells. *Science* **322,** 1855–7. ISSN: 1095-9203 (Electronic) 0036-8075 (Linking) (2008).

89. Heppner, G. H., Dexter, D. L., DeNucci, T., Miller, F. R. & Calabresi, P. Heterogeneity in drug sensitivity among tumor cell subpopulations of a single mammary tumor. *Cancer research* **38,** 3758–3763 (1978).

90. Hollstein, M., Sidransky, D., Vogelstein, B. & Harris, C. C. p53 mutations in human cancers. *Science* **253,** 49–54 (1991).

91. Holm, S. A simple sequentially rejective multiple test procedure. *Scandinavian journal of statistics,* 65–70. ISSN: 0303-6898 (1979).

92. Van Houwelingen, H. C., Bruinsma, T., Hart, A. A., van't Veer, L. J. & Wessels, L. F. Cross-validated Cox regression on microarray gene expression data. *Statistics in medicine* **25,** 3201–3216. ISSN: 1097-0258 (2006).

93. Hsu, F., Kent, W. J., Clawson, H., Kuhn, R. M., Diekhans, M. & Haussler, D. The UCSC known genes. *Bioinformatics* **22,** 1036–1046. ISSN: 1367-4803 (2006).

94. Hu, P., Lan, H., Xu, W., Beyene, J. & Greenwood, C. M. Identifying cis-and trans-acting single-nucleotide polymorphisms controlling lymphocyte gene expression in humans. *BMC proceedings* **1,** 1 (2007).

95. Huang, R., Jaritz, M., Guenzl, P., Vlatkovic, I., Sommer, A., Tamir, I. M., Marks, H., Klampfl, T., Kralovics, R. & Stunnenberg, H. G. An RNA-Seq strategy to detect the complete coding and non-coding transcriptome including full-length imprinted macro ncRNAs. *PLoS One* **6,** e27288. ISSN: 1932-6203 (2011).

96. Huang, S., Chong, N., Lewis, N. E., Jia, W., Xie, G. & Garmire, L. X. Novel personalized pathway-based metabolomics models reveal key metabolic pathways for breast cancer diagnosis. *Genome medicine* **8,** 1. ISSN: 1756-994X (2016).

97. Huang, S., Yee, C., Ching, T., Yu, H. & Garmire, L. X. A Novel Model to Combine Clinical and Pathway-Based Transcriptomic Information for the Prognosis Prediction of Breast Cancer. *PLoS computational biology* **10,** e1003851. ISSN: 1553-7358 (2014).

98. Ishwaran, H., Kogalur, U. B., Blackstone, E. H. & Lauer, M. S. Random survival forests. *The Annals of Applied Statistics,* 841–860. ISSN: 1932-6157 (2008).

99. Iyer, M. K., Niknafs, Y. S., Malik, R., Singhal, U., Sahu, A., Hosono, Y., Barrette, T. R., Prensner, J. R., Evans, J. R. & Zhao, S. The landscape of long noncoding RNAs in the human transcriptome. *Nature Genetics.* ISSN: 1061-4036 (2015).

100. Izutsu, T., Konda, R., Sugimura, J., Iwasaki, K. & Fujioka, T. Brain-specific angiogenesis inhibitor 1 is a putative factor for inhibition of neovascular formation in renal cell carcinoma. *The Journal of urology* **185,** 2353–2358. ISSN: 0022-5347 (2011).

101. Jalali, S., Jayaraj, G. G. & Scaria, V. Integrative transcriptome analysis suggest processing of a subset of long non-coding RNAs to small RNAs. *Biol Direct* **7,** 25. ISSN: 1745-6150 (Electronic) 1745-6150 (Linking) (2012).

102. Ji, P., Diederichs, S., Wang, W., Boeing, S., Metzger, R., Schneider, P. M., Tidow, N., Brandt, B., Buerger, H. & Bulk, E. MALAT-1, a novel noncoding RNA, and thymosin beta-4 predict metastasis and survival in early-stage non-small cell lung cancer. *Oncogene* **22,** 8031–8041. ISSN: 0950-9232 (2003).

103. Jiang, H. & Wong, W. H. Statistical inferences for isoform expression in RNA-Seq. *Bioinformatics* **25,** 1026–1032 (2009).

104. Jiang, Q., Wang, J., Wu, X., Ma, R., Zhang, T., Jin, S., Han, Z., Tan, R., Peng, J. & Liu, G. LncRNA2Target: a database for differentially expressed genes after lncRNA knockdown or overexpression. *Nucleic acids research* **43,** D193–D196. ISSN: 0305-1048 (2015).

105. Jones, N. The learning machines. *Nature* (2014).

106. Joshi, R. & Reeves, C. *Beyond the Cox model: artificial neural networks for survival analysis part II* in *Proceedings of the eighteenth international conference on systems engineering* (2006), 179–184.

107. Joyce, G. F. The antiquity of RNA-based evolution. *Nature* **418,** 214–221. ISSN: 0028-0836 (2002).

108. Kandoth, C., McLellan, M. D., Vandin, F., Ye, K., Niu, B., Lu, C., Xie, M., Zhang, Q., McMichael, J. F., Wyczalkowski, M. A. & et al. Mutational landscape and significance across 12 major cancer types. *Nature* **502,** 333–339. ISSN: 1476-4687 (Electronic) 0028-0836 (Linking) (2013).

109. Kanehisa, M. & Goto, S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic acids research* **28,** 27–30. ISSN: 0305-1048 (2000).

110. Kapusta, A. & Feschotte, C. Volatile evolution of long noncoding RNA repertoires: mechanisms and biological implications. *Trends in Genetics* **30,** 439–452. ISSN: 0168-9525 (2014).

111. Kawaji, H., Lizio, M., Itoh, M., Kanamori-Katayama, M., Kaiho, A., Nishiyori-Sueki, H., Shin, J. W., Kojima-Ishiyama, M., Kawano, M. & Murata, M. Comparison of CAGE and RNA-seq transcriptome profiling using clonally amplified and single-molecule next-generation sequencing. *Genome research* **24,** 708–717. ISSN: 1088-9051 (2014).

112. Kelley, D. & Rinn, J. Transposable elements reveal a stem cell-specific class of long noncoding RNAs. *Genome Biology* **13.** ISSN: 1474-7596. doi:Doi10.1186/Gb-2012-13-11-R107 (2012).

113. Khalil, A. M., Guttman, M., Huarte, M., Garber, M., Raj, A., Morales, D. R., Thomas, K., Presser, A., Bernstein, B. E. & van Oudenaarden, A. Many human large intergenic noncoding RNAs associate with chromatin-modifying complexes and affect gene expression. *Proceedings of the National Academy of Sciences* **106,** 11667–11672. ISSN: 0027-8424 (2009).

114. Kim, D. & Salzberg, S. L. TopHat-Fusion: an algorithm for discovery of novel fusion transcripts. *Genome Biology* **12.** ISSN: 1465-6906. doi:10.1186/gb-2011-12-8-r72 (2011).

115. Kong, L., Zhang, Y., Ye, Z. Q., Liu, X. Q., Zhao, S. Q., Wei, L. & Gao, G. CPC: assess the protein-coding potential of transcripts using sequence features and support vector machine. *Nucleic Acids Res* **35,** W345–9. ISSN: 1362-4962 (Electronic) 0305-1048 (Linking) (2007).

116. Kowalczyk, M. S., Higgs, D. R. & Gingeras, T. R. Molecular biology: RNA discrimination. *Nature* **482,** 310–311. ISSN: 0028-0836 (2012).

117. Koziol, J. A. & Jia, Z. The concordance index C and the Mann-Whitney parameter Pr (X¿ Y) with randomly censored data. *Biometrical Journal* **51,** 467–474. ISSN: 1521-4036 (2009).

118. Kruger, J. & Rehmsmeier, M. RNAhybrid: microRNA target prediction easy, fast and flexible. *Nucleic acids research* **34,** W451–W454. ISSN: 0305-1048 (2006).

119. Kudo, S., Konda, R., Obara, W., Kudo, D., Tani, K., Nakamura, Y. & Fujioka, T. Inhibition of tumor growth through suppression of angiogenesis by brain-specific angiogenesis inhibitor 1 gene transfer in murine renal cell carcinoma. *Oncology reports* **18,** 785–792. ISSN: 1021-335X (2007).

120. Kumar, M., Gromiha, M. M. & Raghava, G. P. SVM based prediction of RNA-binding proteins using binding residues and evolutionary information. *Journal of Molecular Recognition* **24,** 303–313. ISSN: 1099-1352 (2011).

121. Kumar, V., Westra, H.-J., Karjalainen, J., Zhernakova, D. V., Esko, T., Hrdlickova, B., Almeida, R., Zhernakova, A., Reinmaa, E., Võsa, U., *et al.* Human disease-associated genetic variation impacts large intergenic non-coding RNA expression. *PLoS Genet* **9,** e1003201 (2013).

122. Kvam, V. M., Liu, P. & Si, Y. A comparison of statistical methods for detecting differentially expressed genes from RNA-seq data. *American journal of botany* **99,** 248–256 (2012).

123. Labialle, S. & Cavaille, J. Do repeated arrays of regulatory small-RNA genes elicit genomic imprinting? *Bioessays* **33,** 565–573. ISSN: 1521-1878 (2011).

124. LeCun, Y. & Bengio, Y. Convolutional networks for images, speech, and time series. *The handbook of brain theory and neural networks* **3361,** 1995 (1995).

125. Lee, J., Koh, J., Shin, B., Ahn, K., Roh, J., Kim, Y. & Kim, K. K. Comparative study of angiostatic and anti-invasive gene expressions as prognostic factors in gastric cancer. *International journal of oncology* **18,** 355–362. ISSN: 1019-6439 (2001).

126. Lee, W., Jiang, Z., Liu, J., Haverty, P. M., Guan, Y., Stinson, J., Yue, P., Zhang, Y., Pant, K. P., Bhatt, D., *et al.* The mutation spectrum revealed by paired genome sequences from a lung cancer patient. *Nature* **465,** 473–477 (2010).

127. Leng, N., Dawson, J. A., Thomson, J. A., Ruotti, V., Rissman, A. I., Smits, B. M., Haag, J. D., Gould, M. N., Stewart, R. M. & Kendziorski, C. EBSeq: an empirical bayes hierarchical model for inference in RNA-seq experiments. *Bioinformatics* **29,** 1035–1043. ISSN: 1367-4803 (2013).

128. Lewis, B. P., Burge, C. B. & Bartel, D. P. Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. *cell* **120,** 15–20. ISSN: 0092-8674 (2005).

129. Li, C. I., Su, P. F. & Shyr, Y. Sample size calculation based on exact test for assessing differential expression analysis in RNA-seq data. *BMC bioinformatics* **14,** 357. ISSN: 1471-2105 (Electronic) 1471-2105 (Linking) (2013).

130. Li, J., Poursat, M.-A., Drubay, D., Motz, A., Saci, Z., Morillon, A., Michiels, S. & Gautheret, D. A dual model for prioritizing cancer mutations in the non-coding genome based on germline and somatic events. *PLoS Comput Biol* **11,** e1004583 (2015).

131. Liang, C. C., Park, A. Y. & Guan, J. L. In vitro scratch assay: a convenient and inexpensive method for analysis of cell migration in vitro. *Nat Protoc* **2,** 329–33. ISSN: 1750-2799 (Electronic) 1750-2799 (Linking) (2007).

132. Liao, Q., Liu, C., Yuan, X., Kang, S., Miao, R., Xiao, H., Zhao, G., Luo, H., Bu, D. & Zhao, H. Large-scale prediction of long non-coding RNA functions in a coding–non-coding gene co-expression network. *Nucleic acids research* **39,** 3864–3878. ISSN: 0305-1048 (2011).

133. Liao, Y., Smyth, G. & Shi, W. featureCounts: an efficient general-purpose read summarization program. *arXiv* **1305,** 16 (2013).

134. Liao, Y., Smyth, G. K. & Shi, W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **30,** 923–930. ISSN: 1367-4803 (2014).

135. Ling, H., Spizzo, R., Atlasi, Y., Nicoloso, M., Shimizu, M., Redis, R. S., Nishida, N., GafA, R., Song, J. & Guo, Z. CCAT2, a novel noncoding RNA mapping to 8q24, underlies metastatic progression and chromosomal instability in colon cancer. *Genome research* **23,** 1446–1461. ISSN: 1088-9051 (2013).

136. Liu, H., Yue, D., Chen, Y., Gao, S.-J. & Huang, Y. Improving performance of mammalian microRNA target prediction. *BMC bioinformatics* **11,** 476. ISSN: 1471-2105 (2010).

137. Liu, K., Yan, Z., Li, Y. & Sun, Z. Linc2GO: a human LincRNA function annotation resource based on ceRNA hypothesis. *Bioinformatics* **29,** 2221–2222. ISSN: 1367-4803 (2013).

138. Liu, Y., Zhou, J. & White, K. P. RNA-seq differential expression studies: more sequence or more replication? *Bioinformatics* **30,** 301–4. ISSN: 1367-4811 (Electronic) 1367-4803 (Linking) (2014).

139. Livak, K. J. & Schmittgen, T. D. Analysis of relative gene expression data using real-time quantitative PCR and the 2-delta-delta-CT method. *methods* **25,** 402–408. ISSN: 1046-2023 (2001).

140. Loewen, G., Zhuo, Y., Zhuang, Y., Jayawickramarajah, J. & Shan, B. lincRNA HOTAIR as a novel promoter of cancer progression. *Journal of Cancer Research Updates* **3,** 134–140. ISSN: 1929-2279 (2014).

141. Love, M., Anders, S. & Huber, W. Differential analysis of RNA-Seq data at the gene level using the DESeq2 package. *Bioconductor* (2013).

142. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome biology* **15,** 1 (2014).

143. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-Seq data with DESeq2. *bioRxiv* (2014).

144. Ma, H., Hao, Y., Dong, X., Gong, Q., Chen, J., Zhang, J. & Tian, W. Molecular mechanisms and function prediction of long noncoding RNA. *The Scientific World Journal* **2012** (2012).

145. Ma, L., Bajic, V. B. & Zhang, Z. On the classification of long non-coding RNAs. *RNA biology* **10,** 924–933 (2013).

146. Ma, W., Ay, F., Lee, C., Gulsoy, G., Deng, X., Cook, S., Hesson, J., Cavanaugh, C., Ware, C. B. & Krumm, A. Fine-scale chromatin interaction maps reveal the cis-regulatory landscape of human lincRNA genes. *Nature methods.* ISSN: 1548-7091 (2014).

147. Maaten, L. v. d. & Hinton, G. Visualizing data using t-SNE. *Journal of Machine Learning Research* **9,** 2579–2605 (2008).

148. Marioni, J. C., Mason, C. E., Mane, S. M., Stephens, M. & Gilad, Y. RNA-seq: An assessment of technical reproducibility and comparison with gene expression arrays. *Genome Research* **18,** 1509–1517. ISSN: 1088-9051, 1549-5469 (2008).

149. Marques, A. C. & Ponting, C. P. Catalogues of mammalian long noncoding RNAs: modest conservation and incompleteness. *Genome Biol* **10,** R124 (2009).

150. Marques, I., Teixeira, A. L., Ferreira, M., Assis, J., Lobo, F., Mauricio, J. & Medeiros, R. Influence of survivin (BIRC5) and caspase-9 (CASP9) functional polymorphisms in renal cell carcinoma development: a study in a southern European population. *Molecular biology reports* **40,** 4819–4826. ISSN: 0301-4851 (2013).

151. Masters, T. *Practical neural network recipes in C++* ISBN: 0124790402 (Morgan Kaufmann, 1993).

152. Mattick, J. S. RNA regulation: a new genetics? *Nature Reviews Genetics* **5,** 316–323 (2004).

153. McCarthy, D. J., Chen, Y. & Smyth, G. K. Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. *Nucleic Acids Research* **40,** 4288–4297. ISSN: 0305-1048, 1362-4962 (2012).

154. McCulloch, C. E. Maximum Likelihood Algorithms for Generalized Linear Mixed Models. *Journal of the American Statistical Association* **92.** ISSN: 01621459. doi:10.2307/2291460 (1997).

155. McCulloch, W. S. & Pitts, W. A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics* **5,** 115–133. ISSN: 0007-4985 (1943).

156. McHugh, C. A., Russell, P. & Guttman, M. Methods for comprehensive experimental identification of RNAa-protein interactions. *Genome Biol* **15,** 203 (2014).

157. McIntyre, L. M., Lopiano, K. K., Morse, A. M., Amin, V., Oberg, A. L., Young, L. J. & Nuzhdin, S. V. RNA-seq: technical variability and sampling. *BMC genomics* **12** (2011).

158. McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M. & et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome research* **20,** 1297–1303 (2010).

159. Menor, M., Ching, T., Zhu, X., Garmire, D. & Garmire, L. X. mirMark: a site-level and UTR-level classifier for miRNA target prediction. *Genome biology* **15,** 500. ISSN: 1465-6906 (2014).

160. Mercer, T. R., Dinger, M. E. & Mattick, J. S. Long non-coding RNAs: insights into functions. *Nature Reviews Genetics* **10,** 155–159. ISSN: 1471-0056 (2009).

161. Mercer, T. R., Gerhardt, D. J., Dinger, M. E., Crawford, J., Trapnell, C., Jeddeloh, J. A., Mattick, J. S. & Rinn, J. L. Targeted RNA sequencing reveals the deep complexity of the human transcriptome. *Nat Biotechnol* **30,** 99–104. ISSN: 1546-1696 (Electronic) 1087-0156 (Linking) (2012).

162. Meric-Bernstam, F., Johnson, A., Holla, V., Bailey, A. M., Brusco, L., Chen, K., Routbort, M., Patel, K. P., Zeng, J., Kopetz, S. & et al. A Decision Support Framework for Genomically Informed Investigational Cancer Therapy. *Journal of the National Cancer Institute* **107,** djv098. ISSN: 0027-8874, 1460-2105 (July 2015).

163. Montgomery, S. B., Sammeth, M., Gutierrez-Arcelus, M., Lach, R. P., Ingle, C., Nisbett, J., Guigo, R. & Dermitzakis, E. T. Transcriptome genetics using second generation sequencing in a Caucasian population. *Nature* **464,** 773–7. ISSN: 0028-0836 (2010).

164. Morin, R. D., Bainbridge, M., Fejes, A., Hirst, M., Krzywinski, M., Pugh, T. J., McDonald, H., Varhol, R., Jones, S. J. M. & Marra, M. A. Profiling the HeLa S3 transcriptome using randomly primed cDNA and massively parallel short-read sequencing. *Biotechniques* **45** (2008).

165. Muppirala, U., Lewis, B. A. & Dobbs, D. Computational tools for investigating RNA-protein interaction partners. *J Comput Sci Syst Biol* **6,** 182–187 (2013).

166. Murphy, K. & Mian, S. *Modelling gene expression data using dynamic Bayesian networks* Report (Technical report, Computer Science Division, University of California, Berkeley, CA, 1999).

167. Mutch, D. M., Berger, A., Mansourian, R., Rytz, A. & Roberts, M.-A. The limit fold change model: a practical approach for selecting differentially expressed genes from microarray data. *BMC bioinformatics* **3,** 17. ISSN: 1471-2105 (2002).

168. Ng, S. B., Turner, E. H., Robertson, P. D., Flygare, S. D., Bigham, A. W., Lee, C., Shaffer, T., Wong, M., Bhattacharjee, A., Eichler, E. E. & et al. Targeted Capture and Massively Parallel Sequencing of Twelve Human Exomes. *Nature* **461,** 272–276. ISSN: 0028-0836 (Sept. 2009).

169. NHGRI. *An Overview of the Human Genome Project* 2017.

170. Ning, S., Wang, P., Ye, J., Li, X., Li, R., Zhao, Z., Huo, X., Wang, L., Li, F. & Li, X. A global map for dissecting phenotypic variants in human lincRNAs. *European Journal of Human Genetics* **21,** 1128–1133. ISSN: 1018-4813 (2013).

171. Ning, S., Zhang, J., Wang, P., Zhi, H., Wang, J., Liu, Y., Gao, Y., Guo, M., Yue, M., Wang, L. & et al. Lnc2Cancer: a manually curated database of experimentally supported lncRNAs associated with various human cancers. *Nucleic acids research,* gkv1094 (2015).

172. Nishimori, H., Shiratsuchi, T., Urano, T., Kimura, Y., Kiyono, K., Tatsumi, K., Yoshida, S., Ono, M., Kuwano, M. & Nakamura, Y. A novel brain-specific p53-target gene, BAI1, containing thrombospondin type 1 repeats inhibits experimental angiogenesis. *Oncogene* **15,** 2145–2150. ISSN: 0950-9232 (1997).

173. Nookaew, I., Papini, M., Pornputtapong, N., Scalcinati, G., Fagerberg, L., UhlA(C)n, M. & Nielsen, J. A comprehensive comparison of RNA-Seq-based transcriptome analysis from reads to differential gene expression and cross-comparison with microarrays: a case study in Saccharomyces cerevisiae. *Nucleic Acids Research* **40,** 10084–10097 (2012).

174. O'Rawe, J., Jiang, T., Sun, G., Wu, Y., Wang, W., Hu, J., Bodily, P., Tian, L., Hakonarson, H., Johnson, W. E. & et al. Low concordance of multiple variant-calling pipelines: practical implications for exome and genome sequencing. *Genome medicine* **5,** 1 (2013).

175. Oka, H., Chatani, Y., Hoshino, R., Ogawa, O., Kakehi, Y., Terachi, T., Okada, Y., Kawaichi, M., Kohno, M. & Yoshida, O. Constitutive activation of mitogen-activated protein (MAP) kinases in human renal cell carcinoma. *Cancer research* **55,** 4182–4187. ISSN: 0008-5472 (1995).

176. Olden, J. D., Joy, M. K. & Death, R. G. An accurate comparison of methods for quantifying variable importance in artificial neural networks using simulated data. *Ecological Modelling* **178,** 389–397. ISSN: 0304-3800 (2004).

177. Parsell, M. Steven M. Platek, Julian Paul Keenan and Todd K. Shackelford (eds), Evolutionary Cognitive Neuroscience. *Minds and Machines* **19,** 275–278. ISSN: 0924-6495 (2009).

178. Peng, Z., Cheng, Y., Tan, B. C.-M., Kang, L., Tian, Z., Zhu, Y., Zhang, W., Liang, Y., Hu, X., Tan, X. & et al. Comprehensive analysis of RNA-Seq data reveals extensive RNA editing in a human transcriptome. *Nature biotechnology* **30,** 253–260 (2012).

179. Penny, G. D., Kay, G. F., Sheardown, S. A., Rastan, S. & Brockdorff, N. Requirement for Xist in X chromosome inactivation. *Nature* **379,** 131–137. ISSN: 0028-0836 (1996).

180. Petalidis, L. P., Oulas, A., Backlund, M., Wayland, M. T., Liu, L., Plant, K., Happerfield, L., Freeman, T. C., Poirazi, P. & Collins, V. P. Improved grading and survival prediction of human astrocytic brain tumors by artificial neural network analysis of gene expression microarray data. *Molecular cancer therapeutics* **7,** 1013–1024. ISSN: 1535-7163 (2008).

181. Pham, T. V. & Jimenez, C. R. An accurate paired sample test for count data. *Bioinformatics* **28,** i596–i602 (2012).

182. Pickrell, J. K., Marioni, J. C., Pai, A. A., Degner, J. F., Engelhardt, B. E., Nkadori, E., Veyrieras, J.-B., Stephens, M., Gilad, Y. & Pritchard, J. K. Understanding mechanisms underlying human gene expression variation with RNA sequencing. *Nature* **464,** 768–772 (2010).

183. Piskol, R., Ramaswami, G. & Li, J. B. Reliable identification of genomic variants from RNA-seq data. *The American Journal of Human Genetics* **93,** 641–651 (2013).

184. Prensner, J. R. & Chinnaiyan, A. M. The emergence of lncRNAs in cancer biology. *Cancer discovery* **1,** 391–407 (2011).

185. Prensner, J. R., Iyer, M. K., Balbin, O. A., Dhanasekaran, S. M., Cao, Q., Brenner, J. C., Laxman, B., Asangani, I. A., Grasso, C. S. & Kominsky, H. D. Transcriptome sequencing across a prostate cancer cohort identifies PCAT-1, an unannotated lincRNA implicated in disease progression. *Nature biotechnology* **29,** 742–749. ISSN: 1087-0156 (2011).

186. Qian, F., Chung, L., Zheng, W., Bruno, V., Alexander, R. P., Wang, Z., Wang, X., Kurscheid, S., Zhao, H. & Fikrig, E. Identification of genes critical for resistance to infection by West Nile virus using RNA-Seq analysis. *Viruses* **5,** 1664 (2013).

187. Qian, N. On the momentum term in gradient descent learning algorithms. *Neural networks* **12,** 145–151. ISSN: 0893-6080 (1999).

188. Qiu, M. T., Hu, J. W., Yin, R. & Xu, L. Long noncoding RNA: an emerging paradigm of cancer research. *Tumour Biol* **34,** 613–20. ISSN: 1423-0380 (Electronic) 1010-4283 (Linking) (2013).

189. Rapaport, F., Khanin, R., Liang, Y., Pirun, M., Krek, A., Zumbo, P., Mason, C. E., Socci, N. D. & Betel, D. Comprehensive evaluation of differential gene expression analysis methods for RNA-seq data. *Genome Biology* **14,** R95. ISSN: 1465-6906 (2013).

190. Rinn, J. L., Kertesz, M., Wang, J. K., Squazzo, S. L., Xu, X., Brugmann, S. A., Goodnough, L. H., Helms, J. A., Farnham, P. J. & Segal, E. Functional Demarcation of Active and Silent Chromatin Domains in Human HOX Loci by Noncoding RNAs. *Cell* **129,** 1311–1323. ISSN: 0092-8674 (2007).

191. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26,** 139–140 (2010).

192. Robinson, M. D. & Oshlack, A. A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol* **11** (2010).

193. Robinson, M. D. & Smyth, G. K. Moderated statistical tests for assessing differences in tag abundance. *Bioinformatics* **23,** 2881–2887. ISSN: 1367-4803, 1460-2059 (2007).

194. Robles, J. A., Qureshi, S. E., Stephen, S. J., Wilson, S. R., Burden, C. J. & Taylor, J. M. Efficient experimental design and analysis strategies for the detection of differential expression using RNA-Sequencing. *BMC genomics* **13** (2012).

195. Rubie, C., Kempf, K., Hans, J., Su, T., Tilton, B., Georg, T., Brittner, B., Ludwig, B. & Schilling, M. Housekeeping gene variability in normal and cancerous colorectal, pancreatic, esophageal, gastric and hepatic tissues. *Molecular and cellular probes* **19,** 101–109. ISSN: 0890-8508 (2005).

196. Sahin, A. A., Edgerton, M. E., Murray, J. L. & Bondy, M. Copy Number Imbalances between Screen-and Symptom-Detected Breast Cancers and Impact on Disease-Free Survival. *Cancer prevention research* (2011).

197. Sakharkar, M. K., Chow, V. T. & Kangueane, P. Distributions of exons and introns in the human genome. *In Silico Biol* **4,** 387–93. ISSN: 1386-6338 (Print) 1386-6338 (Linking) (2004).

198. Salmena, L., Poliseno, L., Tay, Y., Kats, L. & Pandolfi, P. P. A ceRNA hypothesis: the Rosetta Stone of a hidden RNA language? *Cell* **146,** 353–8. ISSN: 1097-4172 (Electronic) 0092-8674 (Linking) (2011).

199. Sauvageau, M., Goff, L. A., Lodato, S., Bonev, B., Groff, A. F., Gerhardinger, C., Sanchez-Gomez, D. B., Hacisuleyman, E., Li, E. & Spence, M. Multiple knockout mouse models reveal lincRNAs are required for life and brain development. *Elife* **2,** e01749. ISSN: 2050-084X (2013).

200. Schuster-Bockler, B. & Lehner, B. Chromatin organization is a major influence on regional mutation rates in human cancer cells. *nature* **488,** 504–507 (2012).

201. Sebastian, L. Fast Folding and Comparison of RNA Secondary Structures. *Chemical Monthly* (1994).

202. Semon, M. & Duret, L. Evidence that functional transcription units cover at least half of the human genome. *Trends in Genetics* **20,** 229–232. ISSN: 0168-9525 (2004).

203. Sergushichev, A. An algorithm for fast preranked gene set enrichment analysis using cumulative statistic calculation. *bioRxiv,* 060012 (2016).

204. Shi, L., Song, L., Fitzgerald, M., Maurer, K., Bagashev, A. & Sullivan, K. E. Noncoding RNAs and LRRFIP1 regulate TNF expression. *J Immunol* **192,** 3057–67. ISSN: 1550-6606 (Electronic) 0022-1767 (Linking) (2014).

205. Smith, K. S., Yadav, V. K., Pei, S., Pollyea, D. A., Jordan, C. T. & De, S. SomVarIUS: somatic variant identification from unpaired tissue samples. *Bioinformatics,* btv685 (2015).

206. Smith, M. A., Gesell, T., Stadler, P. F. & Mattick, J. S. Widespread purifying selection on RNA structure in mammals. *Nucleic acids research* **41,** 8220–8236. ISSN: 0305-1048 (2013).

207. Soneson, C. & Delorenzi, M. A comparison of methods for differential expression analysis of RNA-seq data. *BMC bioinformatics* **14,** 91. ISSN: 1471-2105 (2013).

208. Srivastava, N., Hinton, G. E., Krizhevsky, A., Sutskever, I. & Salakhutdinov, R. Dropout: a simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research* **15,** 1929–1958 (2014).

209. Srivastava, S. & Chen, L. A two-parameter generalized Poisson model to improve the analysis of RNA-seq data. *Nucleic Acids Research* **38,** e170–e170 (2010).

210. Strehl, A. & Ghosh, J. Cluster ensembles – a knowledge reuse framework for combining multiple partitions. *Journal of machine learning research* **3,** 583–617 (2002).

211. Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., Paulovich, A., Pomeroy, S. L., Golub, T. R. & Lander, E. S. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences* **102,** 15545–15550. ISSN: 0027-8424 (2005).

212. Sun, K., Chen, X., Jiang, P., Song, X., Wang, H. & Sun, H. iSeeRNA: identification of long intergenic non-coding RNA transcripts from transcriptome sequencing data. *BMC genomics* **14,** S7. ISSN: 1471-2164 (2013).

213. Tahira, A. C., Kubrusly, M. S., Faria, M. F., Dazzani, B., Fonseca, R. S., Maracaja-Coutinho, V., Verjovski-Almeida, S., Machado, M. C. & Reis, E. M. Long noncoding intronic RNAs are differentially expressed in primary and metastatic pancreatic cancer. *Mol Cancer* **10,** 141. ISSN: 1476-4598 (Electronic) 1476-4598 (Linking) (2011).

214. Tamborero, D., Gonzalez-Perez, A., Perez-Llamas, C., Deu-Pons, J., Kandoth, C., Reimand, u., Lawrence, M. S., Getz, G., Bader, G. D., Ding, L. & et al. Comprehensive identification of mutational cancer driver genes across 12 tumor types. *Scientific Reports* **3.** ISSN: 2045-2322. doi:10.1038/srep02650 (Oct. 2013).

215. Tang, X., Baheti, S., Shameer, K., Thompson, K. J., Wills, Q., Niu, N., Holcomb, I. N., Boutet, S. C., Ramakrishnan, R., Kachergus, J. M. & et al. The eSNV-detect: a computational system to identify expressed single nucleotide variants from transcriptome sequencing data. *Nucleic acids research,* gku1005 (2014).

216. Tarazona, S., Garcia-Alcalde, F., Dopazo, J.-n., Ferrer, A. & Conesa, A. Differential expression in RNA-seq: a matter of depth. *Genome Research* **21,** 2213–2223 (2011).

217. Therneau, T. M. & Grambsch, P. M. *Modeling survival data: extending the Cox model* ISBN: 1475732945 (Springer Science & Business Media, 2000).

218. Trapnell, C., Pachter, L. & Salzberg, S. L. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* **25,** 1105–1111. ISSN: 1367-4803, 1460-2059 (2009).

219. Trapnell, C., Williams, B. A., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M. J., Salzberg, S. L., Wold, B. J. & Pachter, L. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature biotechnology* **28,** 511–515 (2010).

220. Tripathi, V., Ellis, J. D., Shen, Z., Song, D. Y., Pan, Q., Watt, A. T., Freier, S. M., Bennett, C. F., Sharma, A. & Bubulya, P. A. The nuclear-retained noncoding RNA MALAT1 regulates alternative splicing by modulating SR splicing factor phosphorylation. *Molecular cell* **39,** 925–938. ISSN: 1097-2765 (2010).

221. Tuch, B. B., Laborde, R. R., Xu, X., Gu, J., Chung, C. B., Monighetti, C. K., Stanley, S. J., Olsen, K. D., Kasperbauer, J. L. & Moore, E. J. Tumor transcriptome sequencing reveals allelic expression imbalances associated with copy number alterations. *PLoS One* **5** (2010).

222. Ulitsky, I. & Bartel, D. P. lincRNAs: genomics, evolution, and mechanisms. *Cell* **154,** 26–46. ISSN: 0092-8674 (2013).

223. Ulitsky, I., Shkumatava, A., Jan, C. H., Sive, H. & Bartel, D. P. Conserved function of lincRNAs in vertebrate embryonic development despite rapid sequence evolution. *Cell* **147,** 1537–50. ISSN: 1097-4172 (Electronic) 0092-8674 (Linking) (2011).

224. Vijay, N., Poelstra, J. W., Kunstner, A. & Wolf, J. B. W. Challenges and strategies in transcriptome assembly and differential gene expression quantification. A comprehensive in silico assessment of RNA-seq experiments. *Molecular Ecology* **22,** 620–634. ISSN: 1365-294X (2013).

225. Vitiello, M., Tuccoli, A. & Poliseno, L. Long non-coding RNAs in cancer: implications for personalized therapy. *Cellular Oncology,* 1–12. ISSN: 2211-3428 (2014).

226. Volders, P.-J., Verheggen, K., Menschaert, G., Vandepoele, K., Martens, L., Vandesompele, J. & Mestdagh, P. An update on LNCipedia: a database for annotated human lncRNA sequences. *Nucleic acids research* **43,** D174–D180. ISSN: 0305-1048 (2015).

227. Volinia, S. & Croce, C. M. Prognostic microRNA/mRNA signature from the integrated analysis of patients with invasive breast cancer. *Proc Natl Acad Sci U S A* **110,** 7413–7. ISSN: 1091-6490 (Electronic) 0027-8424 (Linking) (2013).

228. Walker, R., Bond, J. P., Tarone, R. E., Harris, C. C., Makalowski, W., Boguski, M. S. & Greenblatt, M. S. Evolutionary conservation and somatic mutation hotspot maps of p53: correlation with p53 protein structural and functional features. *Oncogene* **18,** 211–218 (1999).

229. Wang, J., Liu, X., Wu, H., Ni, P., Gu, Z., Qiao, Y., Chen, N., Sun, F. & Fan, Q. CREB up-regulates long non-coding RNA, HULC expression through interaction with microRNA-372 in liver cancer. *Nucleic Acids Res* **38,** 5366–83. ISSN: 1362-4962 (Electronic) 0305-1048 (Linking) (2010).

230. Wang, K., Singh, D., Zeng, Z., Coleman, S. J., Huang, Y., Savich, G. L., He, X., Mieczkowski, P., Grimm, S. A. & Perou, C. M. MapSplice: accurate mapping of RNA-seq reads for splice junction discovery. *Nucleic acids research,* gkq622. ISSN: 0305-1048 (2010).

231. Wang, K. C. & Chang, H. Y. Molecular Mechanisms of Long Noncoding RNAs. *Molecular Cell* **43,** 904–914. ISSN: 1097-2765 (2011).

232. Wang, L., Feng, Z., Wang, X., Wang, X. & Zhang, X. DEGseq: an R package for identifying differentially expressed genes from RNA-seq data. *Bioinformatics* **26,** 136–138 (2010).

233. Wang, L., Park, H. J., Dasari, S., Wang, S., Kocher, J.-P. & Li, W. CPAT: Coding-Potential Assessment Tool using an alignment-free logistic regression model. *Nucleic acids research* **41,** e74–e74. ISSN: 0305-1048 (2013).

234. Wang, Y., Li, Y., Wang, Q., Lv, Y., Wang, S., Chen, X., Yu, X., Jiang, W. & Li, X. Computational identification of human long intergenic non-coding RNAs using a GA-SVM algorithm. *Gene* **533,** 94–99. ISSN: 0378-1119 (2014).

235. Wang, Y., Xu, Z., Jiang, J., Xu, C., Kang, J., Xiao, L., Wu, M., Xiong, J., Guo, X. & Liu, H. Endogenous miRNA sponge lincRNA-RoR regulates Oct4, Nanog, and Sox2 in human embryonic stem cell self-renewal. *Dev Cell* **25,** 69–80. ISSN: 1878-1551 (Electronic) 1534-5807 (Linking) (2013).

236. Wang, Z., Gerstein, M. & Snyder, M. RNA-Seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics* **10,** 57–63. ISSN: 1471-0056 (2009).

237. Warr, A., Robert, C., Hume, D., Archibald, A., Deeb, N. & Watson, M. Exome Sequencing: Current and Future Perspectives. *G3: Genes—Genomes—Genetics* **5,** 1543–1550. ISSN: , 2160-1836 (Aug. 2015).

238. Weakley, S. M., Wang, H., Yao, Q. & Chen, C. Expression and function of a large non-coding RNA gene XIST in human cancer. *World journal of surgery* **35,** 1751–1756. ISSN: 0364-2313 (2011).

239. Wei, R., De Vivo, I., Huang, S., Zhu, X., Risch, H., Moore, J. H., Yu, H. & Garmire, L. X. Meta-dimensional data integration identifies critical pathways for susceptibility, tumorigenesis and progression of endometrial cancer. *Oncotarget.* ISSN: 1949-2553 (2016).

240. Weinstein, J. N., Collisson, E. A., Mills, G. B., Shaw, K. R. M., Ozenberger, B. A., Ellrott, K., Shmulevich, I., Sander, C., Stuart, J. M. & Network, C. G. A. R. The cancer genome atlas pan-cancer analysis project. *Nature genetics* **45,** 1113–1120. ISSN: 1061-4036 (2013).

241. Wilks, C., Cline, M. S., Weiler, E., Diehkans, M., Craft, B., Martin, C., Murphy, D., Pierce, H., Black, J. & Nelson, D. The Cancer Genomics Hub (CGHub): overcoming cancer through the power of torrential data. *Database* **2014,** bau093. ISSN: 1758-0463 (2014).

242. Wilusz, J. E., Sunwoo, H. & Spector, D. L. Long noncoding RNAs: functional surprises from the RNA world. *Genes Dev* **23,** 1494–504. ISSN: 1549-5477 (Electronic) 0890-9369 (Linking) (2009).

243. Wright, M. W. A short guide to long non-coding RNA gene nomenclature. *Human Genomics* **8.** ISSN: 1473-9542. doi:Doi10.1186/1479-7364-8-7 (2014).

244. Wu, H., Wang, C. & Wu, Z. A new shrinkage estimator for dispersion improves differential expression detection in RNA-seq data. *Biostatistics* **14,** 232–243. ISSN: 1465-4644 (2013).

245. Xie, C., Yuan, J., Li, H., Li, M., Zhao, G., Bu, D., Zhu, W., Wu, W., Chen, R. & Zhao, Y. NONCODEv4: exploring the world of long non-coding RNA genes. *Nucleic acids research* **42,** D98–D103. ISSN: 0305-1048 (2014).

246. Yang, L., Duff, M. O., Graveley, B. R., Carmichael, G. G. & Chen, L.-L. Genomewide characterization of non-polyadenylated RNAs. *Genome Biol* **12,** R16 (2011).

247. Yang, L., Lin, C., Liu, W., Zhang, J., Ohgi, K. A., Grinstein, J. D., Dorrestein, P. C. & Rosenfeld, M. G. ncRNA- and Pc2 methylation-dependent gene relocation between nuclear structures mediates gene activation programs. *Cell* **147,** 773–88. ISSN: 1097-4172 (Electronic) 0092-8674 (Linking) (2011).

248. Yang, Y., Li, H., Hou, S., Hu, B., Liu, J. & Wang, J. The Noncoding RNA Expression Profile and the Effect of lncRNA AK126698 on Cisplatin Resistance in Non-Small-Cell Lung Cancer Cell. *PLOS ONE* **8,** e65309. ISSN: 1932-6203 (May 2013).

249. Yu, D., Huber, W. & Vitek, O. Shrinkage estimation of dispersion in Negative Binomial models for RNA-seq experiments with small sample size. *Bioinformatics* **29,** 1275–82. ISSN: 1367-4803 (2013).

250. Yu, D., Huber, W. & Vitek, O. sSeq: A Simple and Shrinkage Approach of Differential Expression Analysis for RNA-Seq experiments. *Bioconductor* (2013).

251. Yuan, J., Wu, W., Xie, C., Zhao, G., Zhao, Y. & Chen, R. NPInter v2.0: an updated database of ncRNA interactions. *Nucleic acids research* **42,** D104–D108. ISSN: 0305-1048 (2014).

252. Zarate, R., Boni, V., Bandres, E. & Garcia-Foncillas, J. MiRNAs and LincRNAs: Could They Be Considered as Biomarkers in Colorectal Cancer? *International Journal of Molecular Sciences* **13,** 840–865 (Jan. 2012).

253. Zhou, X., Chen, J. & Tang, W. The molecular mechanism of HOTAIR in tumorigenesis, metastasis, and drug resistance. *Acta Biochimica et Biophysica Sinica* **46,** 1011–1015. ISSN: 1672-9145, 1745-7270 (Dec. 2014).